Essays in productivity and efficiency

Essays in productivity and efficiency

Proefschrift

ter verkrijging van de graad van doctor aan de Katholieke Universiteit Brabant, op gezag van de rector magnificus, prof. dr. F. A. van der Duyn Schouten, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 13 september 2002 om 10.15 uur door

Victoria Shestalova

geboren op 16 september 1967 te Lugansk, Oekraïne.

PROMOTOR: prof. dr. A.J.J. Talman
PROMOTOR: prof. dr. E.N. Wolff
COPROMOTOR: dr. M.H. ten Raa

Acknowledgment

This thesis is a collection of papers that were written during my work at the Department of Econometrics of Tilburg University and participation in the doctoral program of CentER. I take this opportunity to thank to the people from the university who made my stay here a productive and enjoyable experience.

My deepest thanks go to my supervisor Thijs ten Raa, whose advice and inspiration have always been a great value to me. Dear Thijs, thank you very much!

I am also very grateful to my promotors Dolf Talman and Edward Wolff and to the other members of the committee, Shawna Grosskopf, Pierre Mohnen, Sergio Perelman and Valter Sorana, for their interest to my research and for many helpful comments and suggestions.

I feel indebted to my coauthor and friend Misja Mikkers for our joint research on regulation, which gave a raise to the second part of my thesis, and to Peter Bogetoft, discussions with whom influenced this work very much.

I am very grateful to all people who commented on my papers, especially to Bernie and Mike who must have had difficult times correcting and improving my English. I am also grateful to Dima and Marcel for their help.

Many thanks to my closest friends in Tilburg, Masha and Vlad.

Finally I would like to thank to my colleagues at the university and at DTe, friends and family for their support throughout.

Tilburg, July 2002.

Contents

Chapter 1

General introduction

The thesis presents a collection of articles in the area of productivity and efficiency. It consists of two independent parts, which are connected by the common subject of productivity and efficiency.

Part I of the thesis treats the alternative approaches to the measurement of total factor productivity (TFP) growth and its decomposition. I discuss the assumptions underlying the different measures of TFP growth and the conditions under which they are equivalent. The interrelationship between the different measures provides an interpretation of their similarities and dissimilarities. The interpretation of the productivity indices becomes especially important in empirical applications. In this part I offer two examples of such applications. I analyze the sectoral productivity performance in industrialized countries and identify the sources of TFP growth.

Part II approaches the issue of productivity and efficiency from a different perspective, which reflects the recent shift in my research interests towards incentive regulation. The issue of evaluating the productivity and efficiency performance of companies is very important in the context of regulation. With incentive-based regulation becoming more popular, quite a few regulatory offices started to apply different benchmarking techniques (for example, Data Envelopment Analysis) to quantify the differences in productivity of regulated companies. The results of these analyses have been used in designing incentive schemes that would force regulated companies to improve their performance in terms of productivity and efficiency. Here, the buzz-word is yardstick competition. I apply such a scheme to the regulation of network monopolies and address related optimal incentives issues.

1.1 Part I. TFP growth and its sources

The first part of the thesis deals with various methodological aspects of the measurement of total factor productivity growth. I review a few different approaches to the concept of TFP growth, namely Index Numbers, Data Envelopment Analysis (DEA) and Input-Output Analysis, and establish links among them. Furthermore, I consider theoretical models leading to new decompositions of TFP growth, allowing me to identify the sources of productivity growth.

I summarize the contribution of each chapter below.

1.1.1 Contribution of chapter 2: Review of approaches to the measurement of TFP growth

This chapter reviews different approaches to the measurement of TFP growth and interrelates them. The point of departure is the macroeconomic concept of the Solow Residual and I explain its relation to alternative measures of TFP growth, particularly, to those applied with a more micro orientation.

I focus on Index Numbers, Data Envelopment Analysis (DEA) and Input-Output Analysis. The latter two are especially important in the context of this thesis because they provide the basis for the models considered in chapters 3 and 4.

It appears that the treatment of prices represents the main conceptual difference between the DEA and the traditional Index Number approaches. The traditional productivity indices rest on the assumption of competitive pricing. Consequently, observable value shares are used as weights in aggregation.

This is in contrast to DEA, which does not assume prices to be competitive. The corresponding TFP growth measure - the Malmquist index - is based on fundamentals of the economies and employs shadow price information obtained from the linear program that determines the production possibility frontier and the reference point on the frontier for a given observation.

The Input-Output analytic framework allows us to take into account intersectoral linkages and provides a measure of TFP growth that is conceptually close to the macroeconomic Solow residual (based on observable value shares). The incorporation of shadow information obtained from the general equilibrium model yields an alternative measure of TFP growth, which is close to the measure resulting from DEA.

Measures of TFP growth that assume no optimizing behavior allow us to factor in efficiency change. In particular, the Malmquist indices can be decomposed into technical change and efficiency change. The technical change component of the Malmquist indices represents a shift of the production frontier and resembles the Solow residual measure. The efficiency change component reflects movements towards the frontier. This decomposition of the Malmquist index will be elaborated in chapter 4, where I apply both sequential and contemporaneous Malmquist indices. A combination of the two will lead to a further decomposition, identifying the business-cycle component of TFP growth.

Furthermore, the incorporation of information on international trade allows us to separate the so-called terms-of-trade effect on TFP growth. The TFP growth decomposition is thus augmented with a third term reflecting the contribution of international trade. The model considered in chapter 3 applies this ideas.

1.1.2 Contribution of chapter 3: General Equilibrium of International TFP growth rates

Chapter 3 presents a study of the total factor productivity (TFP) performance in three major economies: the US, Japan and Europe. I consider a general equilibrium model of the three economies, linked by international trade. This model is then used to estimate their TFP growth at the sectoral and the aggregate level.

The model is based on the fundamentals of the economies; it employs only data on input-output flows, factor inputs across sectors, endowments of primary inputs, consumption, and trade patterns. Optimal final demand vectors are obtained by proportional expansions of observable final demand vectors, given the constraint on technology, endowments of primary inputs and trade surplus. The expansion of demand is achieved by reallocation of scarce resources across sectors of the economies and improving the pattern of international trade.

All prices are endogenous. They are obtained as shadow prices from the model's linear program and then used to measure TFP growth. TFP growth is evaluated at shadow prices and decomposed into technical change, efficiency change and terms-of-trade effects.

The empirical analysis based on this model produces a technical change effect that is highly correlated with the conventional Solow residual measure based on observable prices. This result lends support to the use of the standard measure of

technological change.

1.1.3 Contribution of chapter 4: Sequential Malmquist indices of productivity growth: an application to OECD industrial activities

Chapter 4 deals with Malmquist indices and their decompositions. It emphasizes the relevance of the correct interpretation of the latter to the understanding of the processes that underlie productivity changes. The point is illustrated with the analysis of the evolution of productivity in a few developed countries over the period of 1970-90.

I apply both the DEA methodology with contemporaneous frontiers and the less standard DEA with sequential frontiers. The associated industrial Malmquist productivity indices are decomposed into technical change and efficiency change terms, which represent the well-known sources of productivity growth, 'technical progress' and 'catching up'.

Sequential DEA implies that the frontier can only shift outward, while in a contemporaneous setting both inward and outward frontier shifts are possible. Most of DEA literature applies the second approach. However, for the industries in which technological regress is unlikely to occur, DEA with sequential frontiers provides a more adequate measure for the contribution of technical changes than standard DEA.

In this chapter I interrelate the alternative Malmquist indices in a unifying framework that provides an interpretation to their difference. The consequent decomposition of TFP growth combines three terms; namely technical progress, catching-up and the business cycle effects.

1.2 Part II. Incentive regulation and productivity performance

In this part I will focus on the regulation of natural monopolies in the utility sector. In many countries the utility sector has already been restructured and deregulated as to introduce competition, at least in activities where it is sustainable. In most European countries the utility sector, which has traditionally been a public monopoly,

has been split vertically into separate segments - production, transportation over the network, and supply (or retail). While production and supply activities are considered to be competitive (at least potentially), the transportation activity operated by regional monopolists remains monopoly business. Therefore, a regulatory body is typically assigned to ensure efficient pricing and performance.

One aspect of the performance of a regulated network company that appears to be important, but is difficult to incorporate in the regulatory framework in practice is the quality of supply. The model that I present in part II deals with this issue. The proposed regulation scheme is shown to achieve the optimal quality of supply, while providing the companies with an incentive to improve their production efficiency.

1.2.1 Contribution of chapter 5: Review of literature on regulation

In chapter 5 I review the main problems that arise in regulation of regional natural monopolies and the corresponding literature on the theory of regulation.

The key issue in the regulation theory is solving informational asymmetry between the regulator and the regulated firms. The history of regulation offers different approaches to deal with it. For example, in the traditional (the so-called 'cost-plus') regulation firms are compensated for their incurred costs, including a return on assets that is set by the regulator. Thus the regulator disallows the firms to charge excessive returns on their investment. Another example would be a more recent scheme, referred to as 'price-cap', in which the regulator caps the revenues of regulated firms to stimulate the firms to cut the costs and therefore improve efficiency of their operation.

We will discuss the incentive properties of different regulatory approaches and their impact on the quality of services provided by regulated firms.

1.2.2 Contribution of chapter 6: The model of yardstick competition of network utilities

Regulated prices of network services, such as the provision of electricity, gas and water, have traditionally been based on the own costs of companies. Recently a few regulatory bodies in Europe started to use regimes that unlink prices from costs. In some of these high-powered incentive schemes (referred to as 'yardstick competi-

tion') price caps are based on the performance of other companies, giving companies strong incentives to reduce their own costs. While these incentives can have a beneficial impact on costs in the short run, they might have an adverse effect on the reliability of services in the long run, at least without proper quality regulation. To curb such undesirable effects, yardstick competition should be augmented with some mechanism regulating quality.

This chapter shows how forms of yardstick competition can be extended as to incorporate the aspect of reliability. In particular, we will demonstrate that a yardstick competition scheme that does not penalize network failures, is suboptimal and leads to underinvestment. In contrast, the socially-optimal outcome can be achieved by introducing penalties for undersupply which are equal to the value of the associated losses perceived by the customers. The potentially external costs of inadequate supply are thus internalized by the companies and hence taken into account in investment decision making. Given that the regulator does not observe the firm's technology, the main problem is to determine prices such that utilities have sufficient expected revenue to cover both the efficient cost and the risk of shortfall, which will be reflected in fines that must be paid occasionally. This problem can be solved by introducing a yardstick competition regime and augmenting it as to incorporate the risk of network failure. Since we assume firms to be capable of achieving the common minimum installation cost, the proposed regulation scheme emerges as first best.

Part I

TFP growth and its sources

Chapter 2

Review of approaches to the measurement of TFP growth

2.1 Introduction

This introductory chapter gives an overview of the different approaches that are adopted in the literature on measuring total factor productivity (TFP) growth. I will touch upon the Index Number Approach commonly applied by macroeconomists, Input-Output Analysis (IO), and Data Envelopment Analysis (DEA). The latter two are especially relevant in the context of this thesis, as they provide the basis for the models considered in chapters 3 and 4. I will discuss links among the approaches and their relations to the macroeconomic concept of Solow Residual.

It appears that the treatment of prices represents the main conceptual difference between the approaches. The traditional productivity indices rest on the assumption of observable prices being competitive: factors are paid according to their marginal products. Consequently, when measuring TFP growth, observed value shares are used as weights in aggregation.

This is in contrast to DEA, which does not assume prices to be competitive. The corresponding TFP growth measure - the Malmquist index - is based on fundamentals of the economies and employs shadow price information obtained from a linear program that determines the production possibility frontier and the reference point on the frontier for a given observation.

The input-output analysis framework allows us to take into account intersectoral linkages and yields a measure of TFP that is conceptually close to the macro-

economic Solow residual (based on observable value shares). The incorporation of shadow information obtained from a general equilibrium model provides an alternative measure for TFP growth, which is close to the measure resulting from DEA.

This chapter will proceed as follows. First, we briefly review a few different measures of productivity growth in section 2.2; and then we establish relationships among them in section 2.3.

2.2 Approaches to measuring TFP growth

2.2.1 Solow Residual

Total factor productivity growth is conventionally defined as the growth of real output not explained by the growth of factor inputs and associated with changes in technology.

Solow (1957) suggested a framework for measuring technical changes in an economy. He considered the aggregate production function of the form $Y = F(K, L, t)$, in which Y, K and L denoted aggregate output, capital and labor, and variable t stood for time.[1] Solow defined technical change as "any kind of shift in the aggregate production function"[2] and proposed a way of segregating shifts of the production function from movements along it.

In Solow's setting, under the assumption that factors are paid according to their marginal products, technical change is measured as the difference between the rate of growth of real output of the economy and the weighted sum of the growth rates of real inputs (capital and labor). That is, TFP growth is defined by formula

$$\widehat{T} = \widehat{Y} - w_L \widehat{L} - w_K \widehat{K} \tag{2.1}$$

in which w_L and w_K constitute the shares of labor and capital in production. Here and below 'hats' denote growth rates of the corresponding variables, for example, $\widehat{Y} = \frac{1}{Y}\frac{dY}{dt}$, and notation $\widehat{T}$ is used for TFP growth.

It is easy to show that under constant returns to scale $\widehat{T} = \frac{1}{F}\frac{\partial F}{\partial t}$, therefore, indeed, $\widehat{T}$ represents a shift of the production function. In the special case of neutral

[1] In the original notations by Solow aggregate output was labeled Q instead of Y.

[2] "...I am using the phrase 'technical cange' as a short-hand expression for *any kind of shift* in the production function. Thus slowdowns, speedups, improvements in the education of the labor force, all sorts of things will appear as 'technical change' " (Solow, 1957, p.312.)

changes (those leaving marginal rates of transformation untouched) in which the aggregate production function is represented by $A(t)f(K,L)$ with $A(t)$ regarded as technical coefficient, the formula for TFP growth reduces to

$$\widehat{T} = \widehat{A} \tag{2.2}$$

leading to the interpretation of technical change as a change in the technical coefficient.

Expression (2.1) has been named the Solow residual and referred to as "the measure of our ignorance", in other words, the part of output growth that cannot be explained by the growth of inputs.

The definition used by Solow operates with real output and input. Since both are not homogeneous, the way of their aggregation becomes crucial. In particular, Griliches and Jorgenson (1967) argued that the separation of the value of transaction into price and quantity is conceptually wrong and leads to errors of measurement of both real output and real input. According to Griliches and Jorgenson, the most important errors arise from incorrect aggregation, namely, from using biased estimates for the implicit rental value of capital and labor services, from incorrect accounting for changes in investment and consumption goods prices, etc. After incorporating all those adjustments into their analysis of the US national product accounts for the twenty-year period following World War II, they concluded that "if real product and real factor input were accurately accounted for, the observed rate of growth of total factor productivity was negligible"[3]. In spite of such a conclusion, the paper by Griliches and Jorgenson did not close the discussion on the measurement and explanation of TFP growth, but rather stimulated it, inspiring research on aggregation methods. The next section will present more detail on this.

2.2.2 Index number approach

The formula for the residual introduced in the previous section (2.1) provides a measure of TFP growth on the level of macro economy and is often used by macroeconomists in their computations of TFP growth. However, as we have already mentioned, inputs and outputs are not homogeneous. Thus, to compute the growth of inputs and outputs one must somehow aggregate the data.

[3] Griliches and Jorgenson (1967) p.250.

Consider multiple inputs (e.g., different types of capital and labor) and outputs. Then formula (2.1) has to be modified by incorporating index numbers, which results in a representation of TFP growth as the difference of output and input quantity indices

$$\widehat{T} = \widehat{Q}(y,p) - \widehat{Q}(x,w). \tag{2.3}$$

Here and below notation Q is used for quantity indices, y and x are column vectors of output and input, and p and w are row vectors of output and input prices correspondingly.

The type of an index depends on the specification of Q. Most commonly used ones are those of Divisia, Törnqvist and Fisher defined below.

Continuous-time Divisia indices

If $\widehat{Q}(y,p)$ and $\widehat{Q}(x,w)$ are Divisia quantity indices, which we denote by $\widehat{Q}^D(y,p)$ and $\widehat{Q}^D(x,w)$,[4] the weights are determined as value shares of the corresponding inputs (or outputs) in the total input (output) value. That is,

$$\widehat{Q}^D(y,p) = \sum_i \alpha_i \widehat{y}_i \tag{2.4}$$

$$\widehat{Q}^D(x,w) = \sum_j \beta_j \widehat{x}_j \tag{2.5}$$

$$\alpha_i = \frac{p_i y_i}{py}, \ \beta_j = \frac{w_j x_j}{wx}, \tag{2.6}$$

where y_i, p_i, x_j, w_j are coordinates of vectors y, p, x, w. Then the corresponding, Divisia-based, definition of TFP growth is expressed as

$$\widehat{T}^D = \widehat{Q}^D(y,p) - \widehat{Q}^D(x,w) = \sum_i \alpha_i \widehat{y}_i - \sum_j \beta_j \widehat{x}_j. \tag{2.7}$$

Griliches and Jorgenson (1967) have shown that under the necessary condition for producer equilibrium (all marginal rates of transformation between pairs of inputs and outputs are equal to the corresponding price ratios) these indices measure shifts in the production function in case of multiple inputs and outputs. Therefore, indeed, $\widehat{T}^D$ represents technical change as defined by Solow (1957), or as we call it, the Solow residual.

[4]Here and below the superscripts attached to $\widehat{Q}$ and $\widehat{T}$ refer to the method of measurement; for example, the upper index D in the above expression refers to 'Divisia'.

Griliches and Jorgenson (1967) have also demonstrated that under CRS, given the fundamental accounting identity $py = wx$, one can derive a dual definition for TFP growth as the difference of the corresponding price indices (Divisia price indices). Following them, we obtain

$$\widehat{T}^D = \sum_j \beta_j \widehat{w}_j - \sum_i \alpha_i \widehat{p}_i = \widehat{P}^D(x,w) - \widehat{P}^D(y,p), \tag{2.8}$$

where notation $\widehat{P}^D$ stands for Divisia price indices of input and output of the economy.

Notice that the latter formula is equivalent to $\sum_j \beta_j \left(\widehat{w}_j - \widehat{P}^D(y,p)\right)$, in which β_j is the share of factor j in production and $\left(\widehat{w}_j - \widehat{P}^D(y,p)\right)$ denotes the growth of the real marginal product of this factor. This representation imputes productivity growth to factors of production, justifying the name for the residual: total factor productivity growth. For example, in the case of the aggregate production function with two inputs labor and capital, as in Solow (1957), TFP growth can be represented as the sum of the growth of productivity of labor and capital, $\widehat{T}^D = \beta_L \widehat{w}_L + \beta_K \widehat{w}_K$.

The dual definition of TFP growth will be used in chapter 3, where we define an alternative measure of TFP growth based on shadow prices.

Törnqvist and Fisher indices

The continuous-time Divisia indices introduced above have to be approximated in practice. Many empirical applications do this by means of the Törnqvist indices. The latter are also known as translog indices, because Diewert (1978) related them to the translog production function.

Given data on inputs, outputs and value shares in periods t and $t+1$, the translog quantity indices, $\widehat{Q}_y^T$ and $\widehat{Q}_x^T$, are expressed as follows

$$\widehat{Q}_y^T = \widehat{Q}^T(y^t, y^{t+1}, \alpha^t, \alpha^{t+1}) = \sum_i \frac{1}{2}(\alpha_i^t + \alpha_i^{t+1})(\ln y_i^{t+1} - \ln y_i^t) \tag{2.9}$$

$$\widehat{Q}_x^T = \widehat{Q}^T(x^t, x^{t+1}, \beta^t, \beta^{t+1}) = \sum_i \frac{1}{2}(\beta_j^t + \beta_j^{t+1})(\ln x_j^{t+1} - \ln x_j^t), \tag{2.10}$$

where the value shares in each time are defined the same way as before, i.e., $\alpha_i^t = \frac{p_i^t y_i^t}{p^t y^t}$, $\beta_j^t = \frac{w_j^t x_j^t}{w^t x^t}$ and similarly for α_i^{t+1}, β_j^{t+1}. The corresponding Törnqvist productivity index, $\widehat{T}^T$, is constructed as the difference of the corresponding output

and input quantity indices, in accordance with (2.3)

$$\widehat{T}^T = \widehat{Q}_y^T - \widehat{Q}_x^T. \tag{2.11}$$

Another commonly used productivity index is the Fisher index advocated by Diewert (1992), who for the first time suggested using this type of indices for measuring productivity growth. In accordance with this index, the rate of TFP growth is expressed as the difference between the rates of growth of the Fisher output index and the Fisher input index

$$\widehat{T}^F = \ln \widehat{Q}_y^F - \ln \widehat{Q}_x^F. \tag{2.12}$$

The latter are constructed as the geometric average of the Laspeyres and Paasche quantity indices. For example, for output we have

$$\widehat{Q}_y^F = \widehat{Q}^F(y^t, y^{t+1}, p^t, p^{t+1}) = \left(\widehat{Q}_y^L \widehat{Q}_y^P\right)^{1/2},$$

where

$$\begin{aligned} \widehat{Q}_y^L &= \frac{p^t y^{t+1}}{p^t y^t} \\ \widehat{Q}_y^P &= \frac{p^{t+1} y^{t+1}}{p^{t+1} y^t}. \end{aligned}$$

Similar expressions can be constructed for input.

As shown by Diewert (1992), Fisher productivity indices are economically justified in the sense that there exists a certain production structure from which they could be derived. The necessary assumptions are that of competitive revenue maximizing and cost minimizing behavior and the underlying technology being described by a certain class of functional forms. In addition to that, Fisher indices are known to have a few desirable features, in particular, they satisfy the so-called factor reversal property, which the Törnqvist index fails. The factor reversal property guarantees a correct decomposition of value change into price and quantity changes, which is very important for a correct measurement of productivity change and preserving the duality between the measure (2.7) based on quantities and (2.8) referring to prices.

Comparing the two indices, Diewert (1992) has shown that although conceptually the Fisher-type productivity index performs the best, in most practical applications in the time-series context both Törnqvist and Fisher indices yield similar numerical values.

2.2.3 Input-Output Analysis and measuring TFP growth

In this section we will turn to the approach to TFP measurement adopted by Input-Output literature. This literature considers an economy as a system of sectors linked by production processes. Therefore, the measure of TFP growth encompasses intersectoral linkages. In particular, intermediate inputs are introduced into consideration.

Let us assume that the economy consists of n sectors, each producing a certain commodity and using other commodities as intermediate inputs. According to the national accounting identity

$$p_j y_j = \sum_i p_i y_{ij} + \sum_k w_k x_{kj}, \tag{2.13}$$

where $i, j = 1, 2, ...n$, y_j is the gross output of sector j, p_j is its price, y_{ij} is the quantity of intermediate input supplied to sector j from sector i at price p_i, x_{kj} is the quantity of primary input k engaged in production in sector j, with the corresponding price w_k. Primary inputs are typically labor and capital and their prices are assumed to be uniform within the economy.

In the Input-Output Analysis framework the rate of sectoral productivity growth, $\widehat{t_j}$, is conventionally defined as the difference of the growth rates of output and inputs. It is derived from (2.13) and expressed as

$$\widehat{t_j} = \widehat{y}_j - (p_j y_j)^{-1} \left[\sum_i p_i y_{ij} \widehat{y}_{ij} + \sum_k w_k x_{kj} \widehat{x}_{kj} \right]. \tag{2.14}$$

Introducing the technical coefficients $a_{ij} = (y_j)^{-1} y_{ij}$, $b_{kj} = (y_j)^{-1} x_{kj}$, we obtain the equivalent expression for total factor productivity growth as a weighted sum of the reductions in technical coefficients

$$\widehat{t_j} = -p_j^{-1} \left[\sum_i p_i \widehat{a}_{ij} + \sum_k w_k \widehat{b}_{kj} \right]. \tag{2.15}$$

Therefore, similarly to the pair of formulae (2.1) and (2.2) considered in section 2.2.1, we now have the pair of equivalent formulae for TFP growth (2.14) and (2.15).

The aggregate TFP growth in the economy is represented as a combination of the sectoral productivity growths. The sectoral rates of TFP growth are aggregated to the level of macro-economy, using the value shares of sectoral gross outputs in the

net output of the economy as the weights (the Domar decomposition), which leads to the expression for TFP growth in the economy. (See chapter 3 for more detail.)

Formula (2.15) presents the so-called direct measure of sectoral TFP growth and does not take into account the fact that the intermediate inputs are produced by the system. However, competitive equilibrium being assumed, the prices of outputs and inputs are linked by the relationship $p_j = \sum_i p_i a_{ij} + wb_j$, $i, j = 1, 2, ..., n$, so that changes in prices of intermediates result in TFP changes. After accounting for this, one can obtain the expression for 'effective rates' of TFP growth, which account for indirect effects as well.[5] (See, e.g. Aulin-Ahmavaara, 1999).

Not only production of intermediate inputs can be taken into account; other extensions treat capital input as a produced means of production (Peterson, 1979, Wolff, 1985), or treat both labor and capital as produced by the economy (Aulin-Ahmavaara, 1999).

2.2.4 DEA and Malmquist index approach

In this section I discuss an approach to the measurement of TFP growth that is mostly used in the operations research and management science literature: Data Envelopment Analysis (DEA).

DEA deals with the problem of multiple inputs or outputs differently. It constructs a production frontier and computes the 'distance' between the observation and the frontier. Total factor productivity growth is expressed in terms of changes of the distances. Below I will introduce the main concepts and definitions that are necessary to relate this approach to the preceding ones, leaving a more extended discussion of DEA and Malmquist indices to chapter 4.

Following Färe et al. (1996), we define the output set at time t as $P^t(x) = \{y : x$ can produce $y\}$, where x and y are vectors of inputs and outputs as before. We assume sets $P^t(x)$ to be closed, bounded, convex, and satisfy strong disposability of inputs and constant returns to scale.

The production technology is represented by the output distance function, which is defined for any pair of vectors of inputs and outputs (x, y) and time t as

$$D_o^t(x, y) = \inf\{\theta : y/\theta \in P^t(x)\} \tag{2.16}$$

[5] Wolff (1985) distinguishes the value share effect and inter-industry effect, along with the sectoral technical change effect.

The output distance function measures the maximum possible proportional expansion of all outputs given the inputs.[6]

The Malmquist productivity index can be defined as a ratio of two distance functions, as suggested by Caves, Christensen and Diewert (1982), or as the geometric mean of two CCD-type[7] Malmquist indices. The latter was proposed by Färe et al. (1989).

In the present section we apply the latter definition, that is introduce the formula for the Malmquist index as follows[8]

$$M_o(x^{t+1}, y^{t+1}, x^t, y^t) = \left[\left(\frac{D_o^t(x^{t+1}, y^{t+1})}{D_o^t(x^t, y^t)} \right) \left(\frac{D_o^{t+1}(x^{t+1}, y^{t+1})}{D_o^{t+1}(x^t, y^t)} \right) \right]^{1/2}. \qquad (2.17)$$

Values of M_o in excess of one indicate an improvement of TFP, values less than one mean a decrease. The corresponding value of TFP growth is represented as

$$\widehat{T}^M = \ln M_o(x^{t+1}, y^{t+1}, x^t, y^t). \qquad (2.18)$$

As we can see the definition of Malmquist indices uses information about distances to the production frontiers. The construction of the frontier at each time requires knowledge of data on all 'production units' (economies, in our case) that belong to the reference set. Therefore, to apply the formula (2.17), it is not enough to know information about the 'production unit' in question. One should have data on inputs and outputs for the whole reference set of economies as well. Complications of the

[6] To compute the distance for some observation (x, y) we have to solve the following problem

$$\begin{aligned} &\inf_{\theta, \lambda \geq 0} \theta \\ \text{s .t. } &-y/\theta + Y^T\lambda \geqslant 0 \\ &x - X^T\lambda \geqslant 0 \end{aligned}$$

in which X and Y are matrices composed of vector columns of inputs and outputs corresponding to our sample of production units (economies). Alternatively we could use an input distance function, which shows the maximum possible proportional contraction of all inputs still to be able to produce the same amount of output. This would lead to the same measure of efficiency, because input and output distance functions are equivalent under the assumption of constant returns to scale (see Färe and Grosskopf, 1996).

[7] CCD stands for Caves, Christensen and Dievert (1982).

[8] This is analogous to the Fisher index, which is the geometric average of the Laspeyres and Paasche indices.

reconciliation of the data for international comparison explains why the Malmquist indices are not very popular among the macroeconomists. Just a few studies applied them so far for evaluating aggregate productivity changes (e.g., Färe et al., 1994, Taskin and Zaim, 1997).

However, Malmquist indices have a number of desirable properties, most important of which is the independence of behavioral assumptions such as profit maximization or cost minimization.

Notice that while the Törnqvist and Fisher indices are defined in terms of values, the Malmquist indices use only primary information on inputs and outputs and do not require input prices or output prices in their computation[9]. The explicit price information is replaced by implicit ('shadow') price information, derived from the shape of the frontier. (See Coelli and Psarada Rao, 2001.)

Another, though related, attractive feature of Malmquist productivity indices is that they can be decomposed into economically meaningful sources of TFP growth: technical change (or shifts of the production frontier) and efficiency change (movements relative to the production frontier). I will elaborate on this decomposition in chapter 4 of the thesis.

2.3 Relations between DEA, IO and index number approaches

After reviewing the approaches to TFP growth measurement in the previous section, we proceed with the analysis of the relationships among the different measures. First, in section 2.3.1 we establish the relation between the conventional productivity indices of TFP growth and the Malmquist index and demonstrate that under certain conditions the former are equivalent to the technical change component of the latter. Then in section (2.3.2) we focus on the relation between the measures used by Input-Output Analysis and DEA.

[9] Although in theory the Malmquist indices work with physical inputs and outputs, some information on prices can still be necessary in practice. For example, to use capital as input, one have to be able to measure capital. Then observed prices are needed to aggregate over different capital goods.

2.3.1 Link between the Residual and the Malmquist index

Notice that the assumption of optimizing behavior underlying the Törnqvist and Fisher indices implies that they measure pure technical change and do not account for production inefficiencies. On the contrary, the Malmquist index does not require this behavioral assumption and incorporates inefficiency in the analysis. In fact, technical change as defined by Solow, which is measured by the conventional indices (those considered in section 2.2.2) and identified with shifts of the production frontier, corresponds to the technical change component of the Malmquist index. The following example illustrates this point.

Example 2.1 *Let us consider the case of one output and neutral technical changes. In this case the technology can be represented by a production function of the form*

$$y^t = A(t)F(x^t) \tag{2.19}$$

and the Solow Residual is equivalent to $\widehat{A}$*, which in the discrete case is expressed as*

$$SR = \ln A(t+1) - \ln A(t) = \ln \frac{A(t+1)}{A(t)}. \tag{2.20}$$

It can be shown that in this special case the technical change component of the Malmquist index is equivalent to (2.20).[10] *In particular, notice that for this production function the output distance function at* t *is as follows*

$$D_o^t(x,y) = \min\{\theta : y/\theta \leq A(t)F(x)\} =$$
$$= \min\{\theta : y/A(t)F(x) \leq \theta\} = \frac{y}{A(t)F(x)}.$$

Substituting this into the formula for the Malmquist index yields

$$M_o(x^{t+1}, y^{t+1}, x^t, y^t) = \frac{y^{t+1}}{F(x^{t+1})} \frac{F(x^t)}{y^t}. \tag{2.21}$$

Since we focus on the technical change component, we can restrict ourselves to the case of no inefficiency, in which output and input are related by (2.19) in each time t*. By substituting (2.19) in the last formula, we obtain the expression for the Malmquist index as follows*

$$M_o(x^{t+1}, y^{t+1}, x^t, y^t) = \frac{A(t+1)}{A(t)}, \tag{2.22}$$

which is equivalent to the Solow measure of technical change (2.20) above.

[10] Färe et al. (1994) provides a similar illustration, but their analysis is limited to the case of a Cobb-Douglas production function.

The observation demonstrated in the above example holds in a more general case of nonneutral technical changes. In this respect two important results have been established in the literature.

First, Caves et al. (1982) have shown that the Malmquist index (2.17) becomes a Törnqvist productivity index (2.11) provided that the distance functions are of translog form with identical second order coefficients, and that the prices are those supporting cost minimization and profit maximization.

Second, Färe and Grosskopf (1992) proved that under the assumption of maximizing behavior the Malmquist index (2.17) is approximately equal to the Fisher productivity index (2.12).

These two general results provide a link between the conventional Törnqvist and Fisher productivity indices and the Malmquist index, and formulate the conditions for their equivalence. In both cases the assumption of the optimizing behavior of producers plays the crucial role. Under this assumption all three indices (Törnqvist, Fisher and Malmquist) represent shifts of the production frontier - or 'technical change' as defined by Solow - leading to the interpretation of the technical change component of the Malmquist index as Solow residual.

2.3.2 Synthesis of Input-Output Analysis and DEA

As we have discussed above, the 'effective rates' constructed within the neoclassical Input-Output framework allow us to take into consideration the changes of productivity which are due to changes of relative prices. The optimizing behavior being assumed, the prices used in computation are observable prices.

Ten Raa and Mohnen (2002) augment the neoclassical measure of TFP growth as follows. They apply the traditional formula of the neoclassical growth accounting, but use the shadow prices obtained from the linear program instead of the observable ones. The obtained measure of TFP is based on fundamentals of the economy, similarly to the Malmquist indices.

The underlying linear program is as follows. Given a Leontief technology, Leontief preferences and endowments, the economy expands the final-demand vector by adjusting the trade pattern and reallocating inputs among the sectors. The optimal outcome represents the potential that a multi-sectoral open economy could

achieve under free trade by changing the allocation of production factors across sectors within the economy. This is in contrast to DEA, where the potential for improvement is determined by cross-sectional or intertemporal benchmarking.

The new measure of TFP growth encompasses not only the technical change effect (or Solow Residual), but also the efficiency change and terms-of-trade effects. In case of a closed economy the terms-of-trade effect disappears, and the decomposition will reduce to the sum of technical change and efficiency change as before.

There is, however, an important difference between the models. In DEA the available technology is determined by the so-called best practice that is constructed by combining the technologies of the economies in the sample. Consequently, inefficiency is 'technical inefficiency' measured relatively to that best practice. While in the latter model, the available production technology is assumed to be represented by the observed technical coefficients. Inefficiency stems from the suboptimal allocation of production within the system, or from wasting the resources (not employing the endowed primary inputs in production).[11]

Ten Raa and Mohnen (2002) considered a model of a small open economy. Thus the world prices that defined 'the technology' according to which country could export and import goods were exogenous in the model. I will elaborate on this model in chapter 3, in which I consider three large economies trading among each other. The model considered in chapter 3 completely endogenizes prices.

2.4 Conclusion

In this chapter we described several approaches to the measurement of TFP growth rates. We started with the original approach by Solow (1957) and then considered the Index-Number approach, as well as approaches adopted in Input-Output and DEA literature.

We identified the differences and similarities among different methods and summarized the main results from the literature formulating the conditions under which the different methods may provide equivalent (or close) measures for TFP growth. In particular, the condition of optimizing behavior appears to be crucial in this

[11]Strictly speaking, DEA can incorporate other types of inefficiencies as well (for example, non-radial DEA models can account for the presence of a slack). However, we will not discuss those in this particular application, since the standard Malmquist indices based on DEA with constant returns to scale that are typically used for the TFP measurement operate with technical inefficiency.

respect.

The assumption of the optimizing behavior, which lends theoretical support to the conventional Törnqvist or Fisher indices, while not required in the case of Malmquist indices, explains the main conceptual difference between the conventional and the Malmquist indices. This allows the Malmquist indices to incorporate the effect of efficiency change which is neglected by the other indices.

Input-Output framework provides indices of technical changes conceptually close to the conventional Solow Residual. However, they can be augmented to factor in both efficiency change and the terms-of-trade effect. This can be done if the observable prices are replaced by shadow prices obtained from the optimization problem. Although, similarly to DEA, the efficiency is interpreted as the potential for boosting the production to reach the production possibility frontier, there is an important difference in the meaning of the frontier in the two models. In DEA the potential is determined by the observable best practice (possibly achieved by the other market participants), while in the augmented input-output model it comes from improving allocations of production factors within a multi-sectoral economy.

Chapter 3

General Equilibrium Analysis of International TFP Growth Rates

This chapter[1] elaborates on the model by ten Raa and Mohnen (2002) discussed in section 2.3.2.

I consider a general equilibrium model of three large economies - the US, Japan and Europe - linked by international trade. The model is based on the fundamentals of the economies and employs only data on input-output flows, factor inputs across sectors, consumption, trade patterns and endowments. Prices are endogenous in the model. They are obtained as shadow prices from the model's linear program and then used to measure TFP growth.

Similarly to the paper by ten Raa and Mohnen, TFP growth is evaluated at shadow prices and decomposed into technical change, efficiency change and the terms-of-trade effect. The important distinction, however, lies in the treatment of world prices. In ten Raa and Mohnen (2002), which considers a small open economy, only internal prices were determined endogenously, while international prices were exogenous. In my setting, trade between large economies is considered, world prices become endogenous as well.

The model is applied to analyze the total factor productivity (TFP) performance in the US, Japan and Europe between 1985 and 1990. The new technical change measure will be shown to be highly correlated with the conventional Solow residual, lending support to the latter measure of technical change.

[1]This chapter is based on Shestalova (2001).

3.1 Introduction

The standard of living of the citizens in a national economy may rise for three reasons. First, and foremost, technical progress allows the production of more by less. Secondly, an increase of production efficiency enhances a better use of the available resources. Thirdly, an open economy may benefit from changes in the terms of trade.

The first source of growth, technical progress, is measured by the well-known Solow residual. The second is efficiency change. It shows how much an economy can gain by simply a better allocation of scarce resources across sectors and adjusting its patterns of production and trade accordingly. Some changes in the production pattern may appear to be economic from a resource saving point of view and, therefore, boost productivity. For example, the shift towards electronics not only adds more weight to that sector in the Solow residual, but also facilitates a more efficient use of resource inputs.

Changes in the terms of trade are known to be equivalent to technical progress in theory. In practice, however, few studies ascribe productivity growth to this trade component. Moreover, in a general equilibrium framework encompassing the entire economic system, terms-of-trade changes ought to be reduced to technology and preference shifts, possibly in a partner economy.

In this chapter I measure total factor productivity (TFP) growth in three national economies - namely USA, Japan and Europe - linked by trade. TFP growth comprises three terms: Solow residual, efficiency change and terms-of-trade effect, following ten Raa and Mohnen (2002). In their analysis the terms of trade are exogenous, which is plausible for a small open economy. Since here we are interested in the TFP growth of the main world players it is more appropriate to consider the terms-of-trade effect as endogenous, driven by technology and preference shifts. Roughly speaking, the terms-of-trade effect favors TFP growth of a national economy if the imports become cheaper relative to exports. A clean measurement of technology and preference-shift effects by means of national TFP growth rates requires that inputs and outputs are valuated competitively, for the same reason as exposed by Solow (1957) for a national macro-economy. As observed economies are not perfectly competitive, market prices cannot be used at face value and, therefore, are replaced by endogenous shadow prices to evaluate the TFP growth.

The expression for the Solow residual component in the decomposition of TFP

growth considered in this chapter is similar to that for the conventional measure of TFP growth. However, the former is evaluated at the optimal output levels and shadow prices, while the latter employs the observable prices and output levels.

It can be demonstrated that the conventional measure of TFP growth can be represented as a weighted sum of changes in technical coefficients (see, e.g., Wolff, 1994). The Solow residual component precisely corresponds to this representation and captures the effect of technological changes. It measures the growth of output not attributed to the growth of inputs. Thus, this component accounts for the growth of quantity produced, rather than the changes in the value assigned to these units. If an economy is not divided into sectors, then there are no changes in the real value of a unit of output, the real price does not change and the growth of the quantity produced is the same as the growth of its real value. However, if the economy comprises more sectors, then the growth in terms of the number of units no longer coincides with the growth in terms of value assigned to them. Changes in relative prices cause changes in the real value of a unit of one commodity relatively to that of others and, therefore, changes in the productivity of factors producing it.

Starting with the neoclassical definition of TFP as the difference between the growths of real output and input, and accounting for the effect of relative price changes properly, it will be shown here that the Solow residual is augmented with two additional terms: the efficiency change and the terms-of-trade effect. The efficiency change reveals the change in the gap between the optimal outcome (resulting from the general equilibrium model) and the outcome actually achieved, while the terms-of-trade effect is ascribed to changes in terms of trade. For a closed economy changes in relative prices can be ascribed to changes in the real fundamentals of the economy[2], but for an open economy changes in domestic relative prices may result from changes in the fundamentals of the other economies as well. International trade is the transmitting mechanism, and world prices must be determined to capture these effects.

This chapter proceeds as follows. I introduce a model of a system of economies linked by international trade in section 3.2. The model allows me to determine the competitive levels of production and final demand together with the supporting (shadow) prices, which will be used to compute TFP growth. (See ten Raa and

[2]This is similar to the idea explored by Aulin-Ahmavaara (1999), who shows that fully effective rates of productivity growth can be based solely on the technological characteristics of the production system.

Mohnen, 1998, for the connection between competition and optimization.) Section 3.3 presents a formula for the decomposition of TFP growth. TFP growth is decomposed into three effects. The first corresponds to the conventional Solow residual and reflects the growth due to technological changes. The second is associated with changes in efficiency. And the last term - the terms-of-trade effect - stems from changes in relative prices. Since world relative prices as well as optimal trade patterns are endogenous in the model, the terms-of-trade effect will be fully ascribed to changes in the structures of the economies. Section 3.4 describes the data used to estimate the model and section 3.5 presents the results. The main empirical findings are as follows. First, Solow residuals computed using shadow prices and optimal production levels are highly correlated with those based on the observed prices and output levels. This result lends support to the standard practice of the measurement of Solow residual. Second, I have found that Solow residuals for Europe and the US were lower than those for Japan. In spite of a strong negative terms-of-trade effect, Japan was leading in TFP growth over the period. Section 3.6 summarizes the conclusions.

3.2 The Model

A free trade model of the 'world economy' is applied to find the optimal production and trade patterns, as well as the supporting shadow prices of commodities and factors of production. The 'world' in this model consists of three large economies and 'the rest of the world'. The trade with the rest of the world is pegged at the observed level; consequently, the model describes interactions among the large economies only. World prices corresponding to the optimal activity levels in the considered economies are determined by international trade.

A model of this type has already been used by ten Raa and Mohnen (2001) in their paper on the location of comparative advantages between Canada and Europe. The present chapter extends their model to find the optimal levels of production and the supporting shadow prices for the case of three big economies, namely the United States, Japan and Europe[3], which together cover a significant share of the

[3]The data for Europe were constructed by aggregation of the data for three European economies: Germany, France and the UK.

world trade[4]. I have chosen to aggregate the three European countries into one economy to emphasize the tendency in Europe towards union, leading to closing the existing technological gaps. The fact that trade among three European countries is redefined as intra-trade does not change net export from Europe to any of the other economies.

The model maximizes the level of the world final demand subject to commodity and factor inputs constraints, and given the proportions of the domestic final demand vector in each economy.

Tradable goods are assumed not to be differentiated with respect to a country-producer. The technology of each economy j[5] is described by capital and labor input coefficients k_j, l_j (n-dimensional row vectors) and the commodity input coefficient matrix A_j[6] (an n-dimensional square matrix), where n is the number of different commodities, which is the same as the number of sectors. Capital and labor are mobile across sectors within each economy, but immobile across the economies[7]. The gross output vector of economy j is denoted by x_j (an n-dimensional column vector). The net output of economy j can be expressed as $(I - A_j)x_j$.

Following ten Raa and Mohnen (2001), we assume that consumers have preferences of the Leontief type. This implies that the preferences of consumers in economy j can be described by the vector of domestic final demand of this economy,

[4]The relative sizes of the considered countries in terms of GDP are following: 51% (the US), 21% (Japan), 11% (West Germany), 9% (France) and 8% (the UK). In 1985 the industrialized countries covered about 66% of the total world export, as well as about 68% of the total import (Source: GATT International Trade 1986/87). The five considered countries - the US, Japan, Germany, the UK and France - are the five largest exporters and importers in the world, therefore, their trade constitutes the bulk of these volumes.

[5]Indices 1,2,3 are used for the US, Japan, Europe, respectively.

[6]To define the corresponding technical coefficients the commodity technology model is used. The model assumes that any industry producing a commodity produces it by the same technology, which leads to the expression for the matrix of technical coefficients $A = U(V^T)^{-1}$, where U, V are correspondingly "use" and "make" matrixes. In the traditional one-matrix input-output framework V is assumed to be a diagonal matrix with gross outputs of each sector on the diagonal. Then labor and capital coefficients for each industry are expressed as a ratio of the corresponding factor of production employment in the industry to gross output produced by the industry.

[7]The case of factor mobility across economies can be incorporated by pooling the respective constraints. The case of no factor input mobility across sectors can be incorporated by introducing separate constraints for each sector. The case of differentiated labor (or capital) can be earsily incorporated by introducing constraints on each type of labor (capital).

which is denoted by f_j $(j = 1, 2, 3)$. To maximize utility each economy expands its final demand vector. Final demand includes both consumption and gross investment. The inclusion of investment in the objective function allows us to account for the whole stream of future consumption. Weitzman (1976) demonstrated that for competitive economies domestic final demand measures the present discounted value of future consumption.

The expansion factors for final demands of the three economies are denoted by c_1, c_2 and c_3. We can scan the world production possibility frontier by putting $c_1 = c, c_2 = c\gamma_2$ and $c_3 = c\gamma_3$ and varying γ_2 and γ_3, the direction of expansion. Consequently, the corresponding expanded final demands are cf_1, $c\gamma_2 f_2$ and $c\gamma_3 f_3$. Given weights $(1, \gamma_2, \gamma_3)$ the weighted sum of final demands of the three economies becomes $c(f_1 + \gamma_2 f_2 + \gamma_3 f_3)$. Here c can be interpreted as the expansion factor for the weighted sum of final demands of the three economies.

Each tradable commodity can be consumed as a final good, used in production as an intermediate good or exported. That is $x_j \geqslant A_j x_j + c\gamma_j f_j + \begin{bmatrix} z_j \\ 0 \end{bmatrix}$, $j = 1, 2, 3$. Here $\begin{bmatrix} z_j \\ 0 \end{bmatrix}$ denotes total net export from country j. The commodities are numbered in such a way that nontradable commodities follow tradable commodities. Vector z_j corresponds to tradable commodities. Components of the net export vector that correspond to the nontradable commodities are set to zero.

The vector of total net exports of the three countries with the rest of the world is assumed to be fixed at the observed level. Since the sum of total net exports of the three economies should be at least equal to the total net export from those countries to the rest of the world, we obtain

$$\sum_{j=1}^{3} z_j \geqslant \sum_{j=1}^{3} z_j^0,$$

where z_j^0 corresponds to the observed level of total net export from country j.

The linear program is as follows:

$$\max_{x_j, z_j, c} ce^{\mathrm{T}} \sum_{j=1}^{3} \gamma_j f_j \tag{3.1}$$

subject to:

material balance constraint:

$$(I - A_j)x_j \geqslant c\gamma_j f_j + \begin{bmatrix} z_j \\ 0 \end{bmatrix}, \quad j = 1, 2, 3 \tag{3.2}$$

trade with the rest of the world:

$$\sum_{j=1}^{3} z_j \geqslant \sum_{j=1}^{3} z_j^0 \tag{3.3}$$

factor inputs:

$$k_j x_j \leqslant K_j, \quad l_j x_j \leqslant L_j, \quad j = 1, 2, 3 \tag{3.4}$$

non-negativity:

$$x_j \geqslant 0, \quad j = 1, 2, 3. \tag{3.5}$$

Here γ_1 has been put to one, e^{T} is a unit row vector, T denotes transpose, scalar K_j is the capital stock in country j and scalar L_j is the labor force in country j.

Inequalities (3.2) and (3.3) imply that for any tradable commodity, t, we have a worldwide constraint

$$\sum_{j=1}^{3} \left(\sum_{s=1}^{n} (I_{ts} - A_{ts,j}) \right) x_{s,j} \geqslant \sum_{j=1}^{3} c\gamma_j f_{t,j} + \sum_{j=1}^{3} z_{t,j}^0$$

where the subindexes t, s and ts relate to the corresponding components of vectors and matrices. That is, total net output of the three economies should not be less, than the sum of the three economies' total final demand and the observed total export from the system to the rest of the world. For nontradable commodities the

corresponding components of vector $\begin{bmatrix} z_j \\ 0 \end{bmatrix}$, $j = 1, 2, 3$ are equal to zero, and condition (3.2) implies that each country's final demand for a nontradable commodity, t, cannot exceed the net output of this commodity, or:

$$\left(\sum_{s=1}^{n} (I_{ts} - A_{ts,j}) \right) x_{s,j} \geqslant c\gamma_j f_{t,j}, \quad j = 1, 2, 3.$$

The corresponding dual problem is:

$$\min_{p_{trad}, p_j, r_j, w_j} p_{trad}^{\mathrm{T}} \sum_{j=1}^{3} z_j^0 + \sum_{j=1}^{3} r_j K_j + \sum_{j=1}^{3} w_j L_j \tag{3.6}$$

subject to:

$$-p_j^{\mathrm{T}}(I - A_j) + r_j k_j + w_j l_j - \sigma_j = 0, \qquad j = 1, 2, 3 \tag{3.7}$$

$$p_{trad,j} = p_{trad,} \qquad j = 1, 2, 3 \tag{3.8}$$

$$\sum_{j=1}^{3} p_j^{\mathrm{T}} \gamma_j f_j = e^{\mathrm{T}} \sum_{j=1}^{3} \gamma_j f_j \tag{3.9}$$

$$p_{trad} \geqslant 0, \quad p_j \geqslant 0, \quad w_j \geqslant 0, \quad r_j \geqslant 0, \quad \sigma_j \geqslant 0, \quad j = 1, 2, 3, \tag{3.10}$$

where r_j, w_j are rent and wage rate in country j, σ_j are slacks. Vector $p_j = \begin{bmatrix} p_{trad,j} \\ p_{nontrad,j} \end{bmatrix}$ is a vector of prices in country j. The first block of components, $p_{trad,j}$, corresponds to the tradable commodities and these prices are equalized across countries according to (3.8). Notice that this model does not account for transportation cost, nor tariffs.

The linear program (3.1) - (3.5) basically maximizes the expansion factor c. However, in the objective function the expansion factor c is multiplied by a constant (the value of a weighted sum of final demands). The presence of this constant does not change relative shadow prices of goods and factors, but determines the natural normalization rule for them: the value of the weighted final demand at shadow prices has to be the same as at observable prices. This rule is expressed by condition (3.9) in the dual problem.

A commodity will be produced by a country if and only if the cost of its production does not exceed its price. Therefore, in active sectors the slacks are equal to zero. This reflects the phenomenon of complementary slackness, $\sigma_j x_j = 0$. (See, e.g., ten Raa, 1995.) The complementary slackness condition also gives us $r_j k_j x_j = r_j K_j, w_j l_j x_j = w_j L_j$, $j = 1, 2, 3$.

Multiplying (3.7) by x_j, we obtain that for any country j

$$-p_j^{\mathrm{T}}(I - A_j)x_j + r_j k_j x_j + w_j l_j x_j - \sigma_j x_j = 0.$$

The last expression implies the well-known macroeconomic identity of the national product and national income:

$$p_j^{\mathrm{T}}(I - A_j)x_j = r_j K_j + w_j L_j. \tag{3.11}$$

The condition on net export from the system, (3.3), is binding. Consequently, trade surplus of the system vis-a-vis the rest of the world satisfies:

$$p_{trad}^{\mathrm{T}} \sum_{j=1}^{3} z_j = p_{trad}^{\mathrm{T}} \sum_{j=1}^{3} z_j^0.$$

If we denote the total trade surplus of country j at the optimal point as S_j,

$$S_j = p_{trad}^{\mathrm{T}} z_j,$$

and the trade surplus of country j corresponding to the observable trade pattern as S_j^0,

$$S_j^0 = p_{trad}^{\mathrm{T}} z_j^0,$$

we can express the above condition on total trade with the rest of the world as a condition on the sum of the countries' surpluses in the international trade

$$S_1 + S_2 + S_3 = S_1^0 + S_2^0 + S_3^0. \tag{3.12}$$

The solution of the dual program gives us shadow prices and optimal levels of output in each sector for each economy for the given set of weights γ_2 and γ_3. Thus, we have first to define the weights. We obtain them from the following condition on surpluses of countries in international trade:

$$S_1 = S_1^0, \quad S_2 = S_2^0, \quad S_3 = S_3^0. \tag{3.13}$$

These conditions play the role of a budget constraint on international trade: the obtained equilibrium allocation must preserve the debt positions. At the equilibrium price vector of tradable commodities, p_{trad}, country j can trade the initial quantity z_j^0 for at least S_j^0, but it adjusts its trade, preserving its debt position. By (3.12), any two of the equations (3.13) implies the third one, so they determine the two weights, γ_2 and γ_3, which characterize the optimal welfare distribution among the three economies under free trade. Thus, linear program (3.1) - (3.5) together with condition (3.13) defines an equilibrium level of production and consumption for the three economies.

3.3 The definition of TFP growth

The solution of the above problem provides the optimal allocation of production for a given year, and determines how much the final consumption can be expanded. Hence, the general equilibrium model gives us an economic criterion to define the maximum expansion and the optimal point.

Similarly to Data Envelopment Analysis, we interpret the inverse of the expansion factor of an economy as its efficiency and say that 'the efficiency of economy j' is $(c\gamma_j)^{-1}$. The optimal point represents the state that is feasible to reach under the given assumptions on current technology and preferences. Thus, in accordance with the DEA terminology, we refer to this point as 'the reference point on the frontier'. Consequently, changes of the expansion factor over time are called efficiency changes, while shifts of the optimal point - technical changes. The contribution of these two sources of TFP growth has been acknowledged by DEA literature.

International trade provides another source of TFP growth (see Diewert and Morrison, 1986, ten Raa and Mohnen, 2002). A general equilibrium framework, taking into account international trade, allows us to incorporate this effect.

Following ten Raa and Mohnen (2002), we look at the net import to the economy as an additional input, which together with the traditional inputs - capital and labor

- contribute to the growth of final demand in the economy.

As in Solow (1957) we define the TFP growth as the growth of overall final demand minus the growth of aggregate inputs, however, we use the shadow prices to find the value shares. For any country j we obtain

$$\widehat{TFP}_j = \frac{p^{\mathrm{T}}\dot{f}_j}{p^{\mathrm{T}}f_j} - \frac{w_j\dot{L}_j + r_j\dot{K}_j - p^{\mathrm{T}}_{trad}\dot{z}^o_j}{w_jL_j + r_jK_j - p^{\mathrm{T}}_{trad}z^o_j}, \tag{3.14}$$

in which a dot denotes the time derivative $\frac{d}{dt}$. The subscript j will be dropped in the further derivations to shorten the notation.

The above formula can be rearranged as

$$\widehat{TFP} = \frac{c\gamma p^{\mathrm{T}}\dot{f}}{c\gamma p^{\mathrm{T}}f} - \frac{w\dot{L} + r\dot{K} - p^{\mathrm{T}}_{trad}\dot{z}^o}{wL + rK - p^{\mathrm{T}}_{trad}z^o} =$$

$$= \frac{p^{\mathrm{T}}(c\gamma f)^\bullet - (c\gamma)^\bullet p^{\mathrm{T}}f}{c\gamma p^{\mathrm{T}}f} - \frac{w\dot{L} + r\dot{K} - p^{\mathrm{T}}_{trad}\dot{z}^o}{c\gamma p^{\mathrm{T}}f} =$$

$$= -\frac{(c\gamma)^\bullet}{c\gamma} + \frac{p^{\mathrm{T}}\left(c\gamma f + \begin{bmatrix} z \\ 0 \end{bmatrix}\right)^\bullet - w\dot{L} - r\dot{K} + p^{\mathrm{T}}_{trad}(z^o - z)^\bullet}{c\gamma p^{\mathrm{T}}f} =$$

$$\overset{(3.2)}{=} -\frac{(c\gamma)^\bullet}{c\gamma} + \frac{p^{\mathrm{T}}\left([I - A]\,x\right))^\bullet - r(kx)^\bullet - w(lx)^\bullet}{c\gamma p^{\mathrm{T}}f} + \frac{p^{\mathrm{T}}_{trad}(z^o - z)^\bullet}{c\gamma p^{\mathrm{T}}f} =$$

$$\overset{(3.13)(3.11)}{=} -\frac{(c\gamma)^\bullet}{c\gamma} - \frac{(p^{\mathrm{T}}\dot{A} + r\dot{k} + w\dot{l})x}{c\gamma p^{\mathrm{T}}f} + \frac{\dot{p}^{\mathrm{T}}_{trad}(z - z^o)}{c\gamma p^{\mathrm{T}}f}. \tag{3.15}$$

To derive the above expression we used material balances (3.2), national accounting identities (3.11), and conditions on surpluses (3.13).

Equation (3.15) features three terms. The first term reflects *efficiency change*. Movements of the economy towards the frontier contribute to TFP growth, while outward movements bring about a TFP decline. Hence, TFP is growing when the expansion factor declines.

The second term is *technical change*. As we see it describes the effect of a reduction in technical coefficients for intermediates, capital and labor inputs. In other words, it is the Solow residual evaluated at shadow prices and the optimal gross output levels. Prices enter this term as weights and show the relative importance of technological changes in different sectors.

Even if all technical coefficients and final demand vector in the country remain the same, TFP may still change because of changes in terms of trade, which occur due to shifts in technology or final demand in the other economies. These changes are captured by the last term.

The last term is called the *terms-of-trade effect*, since it is caused by changes in the terms of trade. By (3.15), an increase of the price of a commodity exported in excess of the initially traded quantity yields TFP growth, whilst an increase of the price of an imported commodity leads to a TFP decline. Although we preserve the observed level of the total net export from the system, the terms-of-trade effects for the three economies do not sum up to zero, because of different values of final demand in the denominators.

A similar decomposition of TFP growth has been performed in the paper by ten Raa and Mohnen (2002). In their paper, the observable relative world prices still enter the expression for the TFP growth, because there the case of a small open economy is considered. In the present model the international prices are endogenous (determined by the linear program) and reflect the true marginal cost of production of commodities (at the optimal levels of production and consumption). Therefore, the TFP growth formula relies only on changes in the fundamentals of the economies, namely, endowments, tastes and technologies.

Combining the material balance constraints, (3.2), the condition on trade surpluses, (3.13), and the national account identities, (3.11), we obtain

$$p^{\mathrm{T}} c\gamma f = -p_{trad}^{\mathrm{T}} z^{\mathbf{0}} + rK + wL.$$

Differentiating this condition with respect to time, using (3.14), leads to the dual expression for TFP growth, which imputes the growth of TFP to all factor inputs:

$$\widehat{TFP} = -\frac{\overset{\bullet}{(c\gamma)}}{c\gamma} + \frac{\overset{\bullet}{r}K + \overset{\bullet}{w}L - \overset{\bullet}{p}_{trad}^{\mathrm{T}} z^o}{wL + rK - p_{trad}^{\mathrm{T}} z^o} - \frac{\overset{\bullet}{p}^{\mathrm{T}} f}{p^{\mathrm{T}} f} \tag{3.16}$$

The dual approach to the measurement of TFP growth was first suggested by Jorgenson and Griliches (1967), who showed that under constant returns to scale the

direct definition of the TFP growth as a difference between the growth of quantity of output and quantity of input is equivalent to its dual definition as a difference between the growth of the price of input and the growth of the price of output, or consequently, the growth of real price of input. Formula (3.16) here, however, deviates from that by Jorgenson and Griliches in three respects. First, it incorporates efficiency change - the first term in (3.16). Second, it accounts for international trade and considers net import to the economy as a factor input. Third, it uses shadow prices instead of observable prices.

3.4 Data description

The present analysis is conducted for three economies, namely the US, Japan and Europe, where the latter is an aggregation of France, West Germany and the UK, for the years 1985 and 1990. It uses input-output tables and data on labor and capital stocks across sectors.

The fact that countries use not only different commodity and industry classifications, but also different methodologies of constructing data renders data from national statistical offices incomparable. Reconciliation is a very complicated process requiring additional data at a lower level of aggregation, which is rarely available.

The OECD Statistical Office has made efforts to harmonize the national Input-Output tables of ten OECD countries. The present study makes use of two OECD data bases, namely the Input-Output Data Base (IODB) and the Industrial Structure Data Base (ISDB). The IODB (OECD, 1995) presents the Input-Output tables at several years for ten countries and uses a common industrial classification comprising 36 sectors. The ISDB contains data on the employment and capital stocks. The classification applied in ISDB is less broad (26 sectors, if we exclude subtotals), but can be bridged with the classification used in Input-Output tables. It has to be admitted that the OECD data are still not perfectly harmonized and subject to some inconsistencies, which seem inevitable in the construction of an international data set. However, it is the best alternative available, providing the most complete dataset for the purpose of this research.

The original industrial classification used in IODB distinguishes 35 sectors. Though it is in the interest of this study to have the number of sectors as large as possible, a certain degree of aggregation was required. After the aggregation the number of sectors has been reduced to 31. (For details, see Appendix.)

The input-output tables are converted to constant prices in 1990 US dollars, as follows. First, the tables for 1985[8] are expressed in constant 1990 prices using the ratio of domestic production in constant 1990 prices to domestic production in current 1985 prices in each sector as deflators. Secondly, the tables in constant 1990 prices are converted to constant 1990 dollars by 1990 PPP's. The data on GDP across sectors in current and constant prices for this procedure as well as the PPPs are taken from the ISDB (OECD, 1996). The deflators for observed international prices are constructed as weighted averages of the deflators for the observed domestic prices, with domestic final demands taken as weights in accordance with the normalization used for shadow prices.

Data on labor across sectors comes from ISDB (OECD, 1996) for all countries except for Japan, of which the data is taken directly from the Japan Statistical Yearbook (1995)[9]. Labor is defined as total employment including self-employment and is measured by the number of individuals. Data on labor force for the five countries are taken from the Labor Force Statistics published by OECD (1995). The labor force is given by the number of people who potentially can work.

Data on capital stock by industry comes from ISDB. Capital stocks are estimated by means of the perpetual inventory model. The estimation is based on the series of gross fixed capital formation and specific to each sector and country lives and rates of scrapping (see OECD, 1996 for more detail). For each industry employed capital is defined as the capital stock of industry, corrected for capital utilization. The capital utilization rates[10] for 1985 and 1990 are taken from OECD Economic Outlook (1993)

[8] In fact the table for Germany and the UK presented by OECD are not for 1985 but for 1986 and 1984 correspondingly. It was assumed in this study that the input-output structure in Germany (and the UK) did not change between 1985 and 1986 (and between 1984 and 1985 for the UK). Consequently, the table for Europe was constructed as follows. The tables for Germany (1986) and the UK (1984) were first expressed in constant prices and then added up with the data for France (1985) also expressed in constant prices. The input-output coefficients of the aggregate are weighted sums of the input-output coefficients of each country, the weights being the gross output shares. The OECD tables of 1985 and 1990 for the US are extrapolations of the benchmark table for 1982 using 1977 weights. Updating these tables with more recent information would improve the results.

[9] There were inconsistencies in the data on employment for Japan in ISDB (1996).

[10] We had at our disposal only the average capital utilization rates data for manufacturing. Data on capacity utilization corresponding to agriculture, mining and services were not available. Consequently capacity utilization rates for all industries are assumed to be equal to the average observed for manufacturing.

and The Statistical Abstract of the US (US, Department of Commerce, 1995).

Capital and labor coefficients are constructed as ratios of capital (or labor) employed in the industry to the gross output produced by the industry. For a few sectors with missing values for labor or capital[11], input coefficients are assumed to be equal to the average numbers observed for those sectors in the other countries.

Observed wage and rent shares, which are used for the computation of Solow residuals at the observed prices and output levels, are constructed as follows. Wage shares are obtained from Input-Output tables for all countries except for the US, whose IO Tables do not provide data on compensation of employees. For the US wage shares are taken from the ISDB (OECD, 1996). Rent shares are constructed residually as a difference between value added and wage share.

A bridge table, which links the classification used in the IODB to the classification from ISDB is presented in the Appendix. The ISDB data is slightly more aggregated: some ISDB sectors encompass several IODB sectors. In such cases capital and labor coefficients for each of these IODB sectors are assumed to be equal and computed as a ratio of the capital (or labor) employed in the ISDB sector to the sum of gross outputs produced by the corresponding IODB sectors.

Commodities produced by sectors '23 Construction' and '31 Non-market activities' are considered as nontradable, since the input-output tables for all countries except for Germany reported zero values of export and import for these sectors.

All necessary data for the research have been limited to technical coefficients, endowments of labor and capital and proportions of final demands across sectors. The numerous problems with this kind of international data (e.g., the different treatment of the secondary products by national input-output tables, incomparability of capital utilization rate construction, missing values and absence of cross-sector data on real exchange rate between countries) suggest that the empirical results should be handled with care.

[11] Missing values on capital are encountered in the following sectors: '5 Wood and wood products' (for the US and UK), '11 and 12 Basic metal industries' (the UK), '13 Fabricated metal products' (the UK), '20 Professional goods' (the UK), '21 Other manufacturing industries' (France), '25 Restaurants and hotels' (the US and Japan). Data on labor is missing for sector '25 Restaurants and hotels' for Germany.

3.5 Results of the Total Factor Productivity growth estimation

We start with the analysis of the technical change effect, the first component in equation (3.15). It can be shown that this component can be decomposed further into sectoral technical change effects as follows:

$$SR = \frac{\sum_t p_t x_t SR_t}{c\gamma p^{\mathrm{T}} f}, \tag{3.17}$$

where t stands for sectors, and SR_t denotes the Solow residual in sector t. The last representation shows contributions of each sector t to the total factor productivity growth of the country. The weights used in (3.17) do not sum up to one (the Domar decomposition).

The contribution of sector t to the Solow residual of country j is expressed as

$$SR_{t,j} = \frac{1}{p_{t,j}} \left[-\sum_s p_{s,j} \dot{A}_{st,j} - w_j \dot{l}_{t,j} - r_j \dot{k}_{t,j} \right] = \tag{3.18}$$
$$= -\sum_s \frac{p_{s,j} A_{st,j}}{p_{t,j}} \widehat{A}_{st,j} - \frac{w_j l_{t,j}}{p_{t,j}} \widehat{l}_{t,j} - \frac{r_j k_{t,j}}{p_{t,j}} \widehat{k}_{t,j}.$$

Here a hat denotes the growth rate and subindexes s and t are used for sectors ($s = 1, 2, ..., n; t = 1, 2, ..., n$). For example, $l_{t,j}$ denotes component t of vector l_j. Since the formula above is given for infinitesimal changes, while the computations have to be done for finite changes, we approximate it by the average of two expressions: one with 1985-year weights and one with 1990-year weights. (See e.g. Dietzenbacher and Los, 1998, for discussion of the index number problem.)

The results on sectoral Solow residual are shown in Table 3.1. The first three columns present the Solow residuals of the three economies computed at the observed level of production and using observed prices on commodities and factor inputs. The next three columns correspond to those at the optimum level of production and shadow prices. The sectoral Solow residuals based on observed output and price data are found to be highly correlated with those obtained at the optimum levels of output and using shadow prices. In fact, the correlation coefficients are 0.96 for the US, 0.94 for Japan and 0.92 for Europe.

Table 3.1: Annual Solow residuals (1985-1990)

Industry	SR at observed prices and levels of production (in %)			SR at shadow prices and optimal levels of production (in %)		
	US	**Japan**	**Europe**	**US**	**Japan**	**Europe**
1. Agr., hunting, forestry, fishing	0.84	1.24	2.27	-0.15	3.89	3.19
2. Mining and quarrying	3.20	0.33	1.99	4.61	2.96	4.85
3. Food, beverages, tobacco	-1.03	-1.65	-0.48	-0.97	-1.80	0.01
4. Textiles, wear. apparel, leather	2.32	0.57	0.10	2.02	0.87	1.07
5. Wood and prod., incl. furniture	3.17	2.36	0.06	3.26	2.66	0.71
6. Paper and prod., printing, publ.	-0.65	0.88	-0.09	-0.38	1.33	0.83
7. Ind. chemic., Drugs, medicines	0.43	-0.16	-0.41	0.74	-0.98	-0.84
8. Petroleum and coal	-7.31	1.93	-1.93	-7.31	2.32	-3.05
9. Rubber and plastic products	0.36	-0.54	-0.52	0.56	0.56	-0.75
10. Non-metallic mineral products	1.93	0.08	0.85	2.69	0.52	1.82
11. Iron and steel	-1.86	0.25	-0.82	-2.37	0.36	-0.59
12. Non-ferrous metals	-1.85	0.04	-1.89	-0.50	0.16	-2.08
13. Metal products	-0.60	4.10	1.70	-0.34	5.47	3.33
14. Non-electrical machinery	1.82	2.47	-0.11	2.88	3.58	-0.11
15. Office & computing machinery	5.27	0.90	-2.38	5.98	1.38	-2.78
16. Electric appar., Radio, TV ,etc.	3.61	1.97	0.90	4.37	2.67	1.57
17. Shipbuilding and repairing	0.50	1.09	-0.08	0.69	1.49	1.34
18. Other transport, Motor vehicles	1.28	0.61	0.18	1.72	0.91	1.27
19. Aircrafts	-0.64	1.17	0.70	-0.11	1.25	2.39
20. Professional goods	5.62	1.57	4.53	7.74	2.72	8.12
21. Other manufacturing industries	2.35	-6.15	1.17	4.79	-5.84	1.58
22. Electricity, gas and water	-1.77	0.78	0.62	-1.20	1.61	1.83
23. Construction	-1.71	0.23	0.04	-2.18	1.17	0.86
24. Wholesale and retail trade	3.30	3.88	0.37	3.55	4.69	1.28
25. Restaurants and hotels	0.23	2.37	-1.49	0.58	5.38	-0.78
26. Transport and storage	2.90	2.32	3.10	2.48	3.90	3.69
27. Communication	2.47	1.92	2.04	4.03	3.78	6.94
28. Financ. institut. and insurance	-1.37	-1.11	-2.75	0.38	-0.55	-0.62
29. Real estate, business services	-0.83	0.22	0.30	-1.85	0.29	-2.39
30. Com., soc., personal services	-2.42	-2.79	8.32	-2.43	-2.01	10.85
31. Non-market services	-5.74	-7.93	-4.17	-3.61	-4.99	-3.19
Aggregate SR	-0.26	0.64	0.59	-1.16	3.09	0.03
Aggregate SR (sectors 1-30)[1]	**0.34**	**1.74**	**1.42**	-0.70	4.02	0.93

[1] Sector '31 Non-market services' is not an 'ordinary industry'. First of all, due to general problems of accounting for this kind of services we observe inconsistency in the data reported on them for different countries. Output in non-market services is often measured by their inputs, which makes it impossible to reveal technological changes there. Second, we reallocated all statistical discrepancy to this sector. Inconsistencies in accounting for it as well as the presence of statistical discrepancy may bias the results of the empirical analysis.

The table shows that there is a large difference in Solow residuals between using observed prices and output levels and using shadow prices and optimal output levels. In particular, the latter show a greater variance than the former. This result reflects the fact that optimal prices and production patterns are more volatile than the observed ones. This is because here we assume perfect mobility of labor and capital across sectors and do not account for some trade barriers existing in reality, as for example, the presence of transportation costs. Consequently, the observed prices and output levels differ from those computed in the model. The effect of changes in fundamentals of the economies on the latter is more dramatic than in case of observed prices and production patterns.

The aggregate Solow residuals in the US and Europe were found to be lower than those in Japan for either method. Thus, both methods of computation (with observable prices and outputs and with shadow prices and optimal outputs, respectively) identify Japan as the TFP growth leader over the period 1985-1990.

The results presented here have to be interpreted carefully. To a large extent they are explained by the data used to estimate the model, which themselves inherited some distortions from the original data. For example, the surprisingly low aggregate Solow residual for the US is probably due to the fact that the input-output tables for the US used for the construction of the technical coefficients are not benchmark tables, but just extrapolations (see the footnote on page 36).The results of the TFP growth decomposition (3.15) for the three economies are given in Table 3.2. As explained above the decomposition has been approximated by finite changes, using the average of two decompositions, with weights of 1985 and 1990.

Table 3.2: Decomposition of the annual TFP growth at shadow prices

	Efficiency change	**SR**	**Terms-of-trade effect**	**TFP growth**
US	0.25	-1.16	1.31	0.40
Japan	0.63	3.09	-1.11	2.60
Europe	3.03	0.03	-1.71	1.35

We can see that changes in efficiency were favorable to Europe, while relatively small in the US and Japan. Japan was leading in technical changes over the period.

The terms-of-trade effect was negative in both Japan and Europe, implying that under free trade some of the welfare gains from changes in the fundamentals of the European and Japanese economies would be transmitted towards the other countries.

The aggregate values of TFP growth evaluated at shadow prices are reported in the last column of Table 3.2. In spite of the negative terms-of-trade effect, Japan appeared to be a leader in TFP growth.

The latter finding agrees with that presented in Färe et al. (1994). Färe et al. performed a DEA analysis of TFP growth in OECD countries over 1979-1988 and found that Japan's productivity growth was the highest at the sample. However, the results identifying the sources of TFP growth appear to be different for the two methods: Färe et al. (1994) reports that most of Japan's TFP growth was due to efficiency change, while all of the US growth was due to technical change, which is not in line with what we obtained here.

This urges for a careful interpretation of the results. Notice that although the terminology used here is similar to that used in DEA literature, the way we construct the frontier and measure technical changes and efficiency changes is different from that used in DEA. In DEA the frontier is determined by international benchmarking, while here we evaluate the potential outcome that the system of three economies could have achieved under free trade and perfect mobility of factor inputs. For each economy, the technical-change component, or Solow Residual, represents a weighted sum of reductions in technical coefficients of the economy with weights being based at shadow prices and optimal output levels obtained from the general equilibrium model. This is in contrast to DEA that identifies technical change with TFP growth in the best-practice economies.

Finally, let us compare the values of aggregate TFP growth at shadow prices with those conventionally measured. Notice that the conventional measure of TFP growth is represented by Solow Residuals evaluated at observed prices and output levels. The latter are given in the first three columns of Table 3.1, in which the last two rows show the aggregate numbers for the three economies. The values of SR aggregated over all sectors do not coincide with the values of aggregate TFP growth reported in the last column of Table 3.2. However, the two sets of values become remarkably close after the exclusion of sector '31 Non-market services'. (See the footnote on Table 3.1).

3.6 Conclusion

This chapter introduces a method for the estimation of TFP growth, based solely upon changes in the fundamentals of the economies.

The aggregate TFP growth encompass changes in the marginal valuations of factor inputs in the economies. These marginal valuations result from interactions between all counterparts of the system. Since the economies participating in the system are linked by free trade, changes in tastes, endowments or technologies in any of them affect valuations of inputs in all economies and, therefore, influence TFP growth. Thus we consider international trade as a source of TFP growth in the economies.

TFP growth is evaluated at shadow prices and decomposed into Solow Residual, efficiency change and the effect of change in term of trade. Since the system encompasses three major open economies, terms of trade are endogenous in the model.

The theory has been applied to estimate the TFP growth in the US, Japan and Europe (an aggregate of the UK, France and Germany) in 1985-1990. We have found that the Solow-Residual corresponding to shadow prices and optimal activity levels are strongly correlated with the conventional measure of TFP growth. Japan had the highest aggregate TFP growth over the observed period. This was achieved mostly due to technical change. In contrast, most of the European TFP growth was due to efficiency change.

3.7 Appendix: Bridge table showing the correspondence between IODB and ISDB

	Title of category **IODB**	ISIC code	ISIC code	Title of category **ISDB**
1	Agriculture, hunting, forestry, fishing	1.	1.	Agriculture, hunting, forestry, fishing
2	Mining and quarrying	2.	2.	Mining and quarrying
3	Food, beverages, tobacco	31.	31.	Food, beverages, tobacco
4	Textiles, wearing apparel and leather industries	32.	32.	Textiles, wearing apparel and leather industries
5	Wood and wood products, including furniture	33.	33.	Wood and wood products, including furniture
6	Paper and paper products, printing and publishing	34.	34.	Paper and paper products, printing and publishing
7	Industrial chemicals Drugs and medicines	351.+352.	35.	Chemicals and chemical petroleum, coal, rubber and plastic products
8	Petroleum and coal	353.+354.		
9	Rubber and plastic products	355.+356.		
10	Non-metallic mineral products	36.	36.	Non-metallic mineral products
11	Iron and steel	371.	37.	Basic metal industries
12	Non-ferrous metals	372.		
13	Metal products	381.	381.	Fabricated metal products except machinery and equipment
14	Non-electrical machinery	382.-3825.	382.	Machinery except electrical
15	Office and computing machinery	3825.		
16	Electric apparatus, n.e.c. Radio, TV and communication equipment	383.	383.	Electrical machinery apparatus, appliances and supplies
17	Shipbuilding and repairing	3841.	384.	Transport equipment
18	Other transport Motor vehicles	(384)2.+4.+9. 3843.		
19	Aircrafts	3845.		
20	Professional goods	385.	385.	Professional, scientific, measuring and controlling equipment n.e.c., photographic and optical goods
21	Other manufacturing industries	39.	39.	Other manufacturing industries
22	Electricity, gas and water	4.	4.	Electricity, gas and water
23	Construction	5.	5.	Construction
24	Wholesale and retail trade	61.+62.	61.+62.	Wholesale and retail trade
25	Restaurants and hotels	63.	63.	Restaurants and hotels
26	Transport and storage	71.	71.	Transport and storage
27	Communication	72.	72.	Communication
28	Financial institutions, insurance	81.+82.	81.+82.	Financial institutions, insurance
29	Real estate and business services	83.	83.	Real estate and business services
30	Community, social and personal services	9.	9.	Community, social and personal services
31	Producers of govern. services Other producers Statistical discrepancy			Producers of government services Other producers

Chapter 4

Sequential Malmquist indices of productivity growth: an application to OECD industrial activities

This chapter[1] presents an application of the Malmquist index approach to study the productivity performance in manufacturing industries in a few developed countries over the period 1970-90.

I apply both the standard DEA methodology with contemporaneous frontiers and DEA with sequential frontiers and decompose the associated industrial Malmquist productivity indices into technical change and efficiency change to locate the sources of productivity growth: technical progress and catching up.

The two DEA methodologies differ: sequential DEA implies that the frontier can shift only outward, while in contemporaneous setting both inward and outward frontier shifts are possible. Most of DEA literature applies to the second approach. However, for manufacturing industries, in which technological regress is unlikely to occur, DEA with sequential frontiers provides a more adequate measure for the contribution of technical changes than standard DEA.

Combining the two alternative indices in a unifying framework allows us to distinguish a new component in the Malmquist index decomposition. The new component

[1]The results presented in this chapter were first formulated in Shestalova (2000).

is interpreted as the effect of a business cycle. It reflects shifts in the position of the contemporaneous frontier relatively to the sequential one.

4.1 Introduction

Since the fundamental paper by Solow (1957), in which he paid attention to the unexplained part of the growth of the economy, there were a lot of suggestions in the economic literature on measuring and explaining TFP growth. First, the growth of factors' productivity was viewed purely as a result of technical progress and the fact that an economy may be inefficient was simply neglected. However, later models incorporated efficiency into the analysis and distinguished between two sources of productivity growth: technical progress and catching up. These models construct a production frontier at each point of time and associate technical changes with shifts of the frontier. Changes of the position of observations relative to the frontier are classified as efficiency changes.

The inclusion of inefficiency in the analysis produces changes in the results for TFP growth (as, for example, Färe et al. (1994) have noticed). Moreover, different ways of incorporating inefficiency into the analysis may lead to different estimates for TFP growth or for the components in its decomposition to technical changes and efficiency changes.[2]

We have already discussed some of the possible approaches to the incorporation of inefficiency in chapter 2. One of them is Data Envelopment Analysis - a nonparametric approach that constructs a piecewise linear production frontier by envelopment of available observations on inputs and outputs. In this chapter I apply two types of DEA - contemporaneous and sequential - and analyze the difference between the corresponding indices of TFP growth (contemporaneous and sequential Malmquist indices) and their decompositions. I propose to combine both indices in a common framework, which results in the further decomposition of the Malmquist indices into three components: technical progress, contemporaneous efficiency change and business cycle.

The analysis is applied to the evaluation of productivity performance in manu-

[2]For example, Perelman (1995) compares the outcome of alternative approaches (parametric versus nonparametric) and reports that in his case the discrepancies between the estimates of TFP growth for different approaches are rather satisfactory, whilst the results for the decomposition of TFP growth differ significantly.

facturing industries in OECD countries. There already exist a few studies applying DEA to the international and interregional analysis of productivity performance at the level of industry or economy (see Färe et al. (1994), Perelman (1995), Gouette and Perelman (1997), Taskin and Zaim (1997), Weber and Domazlicky (1999), Cella and Pica (2001), etc.), but they operate with contemporaneous DEA, not with sequential DEA. The contemporaneous DEA assumes that the frontier in each period envelops the observations from this period only. Under such an assumption the technology of previous periods may become unfeasible in the following periods, that is, sometimes the frontier may move inward indicating some 'technical regress'. True, this has a reasonable explanation for industries like mining: the more we have extracted, the more effort and investment it takes to reach deeper layers. But for manufacturing a decline in productivity is usually a temporary phenomenon. Periods of deteriorations alternate with periods of improvement there, which implies that it is unlikely that temporary increases of inputs without increasing output are due to technology deterioration. Classifying these changes as a technological regress may be confusing. In contrast, DEA with sequential frontiers (see, for example, Färe et al. (1985), Tulkens and Vanden Eeckaut (1995)) gives another interpretation to the productivity slowdown. It assumes that in each period of time all preceding technologies are also feasible. The frontier in a certain time envelops all data points observed up to this time, which eliminates the possibility of registering any regress by definition.[3] Another advantage of sequential DEA is practical. Sequential indices incorporate past information and are less sensitive than contemporaneous indices to the presence or not of a particular observation in the sample. I argue, therefore, that sequential DEA provides a more adequate measure of performance than the standard DEA does. In particular, it is more appropriate to use sequential frontiers while evaluating technical changes in manufacturing.

Both contemporaneous and sequential DEA have been applied to the data set covering 6 industries in 11 OECD countries in 1970-1990. I have found that both methods give us highly-correlated measures for the overall TFP growth, but (not surprisingly) less correlated measures for technological changes and for efficiency changes. The correlations between Malmquist indices computed by means of con-

[3]The two cases considered in the present chapter - computations with contemporaneous and sequential frontiers - do not exhaust all possibilities. One can also consider a "window" type of computations (Charnes et al. 1985), in which the frontier in time t is based on a few years of observations.

temporaneous DEA and sequential DEA are above 0.97, whilst the correlations between the technical change components, as well as between efficiency change components, are much lower (ranging from 0.3 to 0.8 across industries).

By splitting the Malmquist indices into three components, I have shown that the discrepancy in the measures of TFP growth that they provide come from the component in their decompositions that represents changes of the position of the contemporaneous frontier relative to the sequential frontier. Two Malmquist indices coincide either if the two frontiers move together, or if shifts of the contemporaneous frontier are Hicks neutral.

The paper is organized as follows. Section 4.2 describes the methodology, which will be used to measure the changes in productivity and efficiency. Section 4.3 presents the data. Section 4.4 discusses empirical results on the Malmquist indices and convergence of productivity, and section 4.5 concludes. The Appendix contains a proof of proposition 4.2.

4.2 Methodology

DEA is a nonparametric method that uses linear programming to construct a nonparametric piecewise frontier of the data. The frontier represents the best practice technology. Observations that belong to it are called efficient by default and the others are inefficient. The efficiency of each observation at a given point in time is measured by means of a distance function, which reflects the distance between the observation and the frontier. The methodology is described in detail in, for example, Färe and Grosskopf (1996).

4.2.1 DEA with contemporaneous frontiers

Let me start with notation. As before, I denote the input and output vectors for one country at time t by $x^t \in \Re_+^n$ and $y^t \in \Re_+^m$, respectively. Let K be the number of countries in our sample. Then $X^t \in \Re_+^{nK}$ and $Y^t \in \Re_+^{mK}$ contain the observations on input and output for all countries in the sample at time t.

Technology in each period t is represented by the output sets $P^t(x) = \{y : x \in \Re_+^n,\ y^t \in \Re_+^m,\ x \text{ can produce } y \text{ in period } t\}$. We assume that the sets $P^t(x)$ satisfy strong disposability of inputs and constant returns to scale. In the *contemporaneous* setting we also assume that any $P^t(x)$ is determined by the observations on inputs

and outputs corresponding to period t, that is,

$$P^t(x) = \{y : y \leqslant Y^t\lambda, x \geqslant X^t\lambda, \lambda \geqslant 0\}, \tag{4.1}$$

where $\lambda \in \Re_+^K$. For any pair of vectors (x, y) we define the output distance function at time t as

$$D_o^t(x, y) = \min\{\theta : y/\theta \in P^t(x)\}. \tag{4.2}$$

The output distance function corresponds to the maximum possible proportional expansion of all outputs given inputs.[4] To compute the distance function for some observation (x, y) we have to solve the following linear program

$$\begin{aligned} &\max_{\eta,\lambda\geq 0}\, \eta \\ \text{s. t. } &-\eta y + Y^t\lambda \geqslant 0 \\ &x - X^t\lambda \geqslant 0, \end{aligned} \tag{4.3}$$

in which $\eta = 1/\theta$. The corresponding value θ will serve as a measure of overall technical efficiency for observation (x, y).

Note that for each observation the distance function reflects the gap between this observation and the frontier, that is, the gap between the observation and the leaders. Closing the gap between leaders and followers implies convergence in total factor productivity. Thus, contemporaneous efficiency introduced above provides us with a natural framework to study the convergence phenomena.

4.2.2 Contemporaneous measure for TFP growth

Färe et al. (1989) suggested using the geometric mean of two CCD-type[5] Malmquist indices to measure TFP growth and to locate its sources. In this chapter we follow the same methodology and consider

[4] Alternatively we could use an input distance function, which shows the maximum possible proportional contraction of all inputs still to be able to produce the same amount of output. This would lead to the same measure of efficiency, because input and output distance functions are equivalent under the assumption of constant returns to scale (see Färe and Grosskopf, 1996).

[5] CCD refers to Caves, Christensen and Diewert (1982), who introduced this type of productivity indices.

$$M_o(x^{t+1}, y^{t+1}, x^t, y^t) = \left[\left(\frac{D_o^t(x^{t+1}, y^{t+1})}{D_o^t(x^t, y^t)} \right) \left(\frac{D_o^{t+1}(x^{t+1}, y^{t+1})}{D_o^{t+1}(x^t, y^t)} \right) \right]^{1/2} \tag{4.4}$$

Rearranging the terms in formula (4.4), following Färe et al. (1989), we obtain the subsequent formula

$$\begin{aligned} & M_o(x^{t+1}, y^{t+1}, x^t, y^t) = \\ & = \frac{D_o^{t+1}(x^{t+1}, y^{t+1})}{D_o^t(x^t, y^t)} \sqrt{\left(\frac{D_o^t(x^t, y^t)}{D_o^{t+1}(x^t, y^t)} \right) \left(\frac{D_o^t(x^{t+1}, y^{t+1})}{D_o^{t+1}(x^{t+1}, y^{t+1})} \right)} \\ & = EFFCH \times TECH. \end{aligned} \tag{4.5}$$

The first factor in equation (4.5) is called efficiency change and shows the change of the relative position of an observation and the frontier. Movements of the observation towards the frontier are associated with values of $EFFCH$ greater than one and are interpreted as efficiency improvements (or 'catching up'). The second factor, the square root term, represents technical change. It corresponds to the shift of the frontier. In particular, outward shifts of the frontier reflect 'technical progress'. An increase in productivity yields a value of the Malmquist index greater than unity and a deterioration leads to a less than unity value. The same holds for each component in the decomposition (4.5) above: any improvement in efficiency or technical progress yields a greater than unity value of the corresponding factor.

Notice that according to the definition, for any time t the contemporaneous frontier envelops the data points of time t and does not depend on data of the previous periods. Under such circumstances the production frontier may shift either inward or outward between t and $t+1$. For manufacturing industries inward shifts of the contemporaneous best practice frontier are usually temporary. Soon the frontier shifts forward, offsetting a deterioration observed earlier. We suggest that shifts of this kind should not qualify as technical change, but as a change of efficiency of the current leaders.

The contribution of technical change can be estimated by means of DEA with sequential frontiers described in detail in Tulkens and Vanden Eeckaut (1995).

4.2.3 DEA with sequential frontiers

Assume that in any period t the technology of the previous period, $t-1$, is still feasible. Consequently, all preceding technologies are feasible as well. Then the production possibility set expands (or remains constant) from one period to the next, the technology can only improve in the course of time, and deteriorations in productivity performance are ascribed to reductions in efficiency.

Generally speaking, the feasibility of the previous period technology would have changed the definition of the output set at time t as follows,

$$\overline{P}^t(x) = \{y : y \leqslant \overline{Y}^t\lambda, x \geqslant \overline{X}^t\lambda, \lambda \geqslant 0\}, \tag{4.6}$$

where $\overline{X}^t = (..., X^{t_0}, ..., X^{t-1}, X^t) = (\overline{X}^{t-1}, X^t)$, $\overline{Y}^t = (..., Y^{t_0}, ..., Y^{t-1}, Y^t) = (\overline{Y}^{t-1}, Y^t)$ and t_0 is the first period, for which observations on inputs and outputs are available. However, the construction of the last set would require information on inputs and outputs before any time t_0. Since this information is missing, we have to truncate set $\overline{P}^t(x)$ at some t_0 and define

$$\overline{P}^t(x|\overline{X}^{t_0} = X^{t_0}, \overline{Y}^{t_0} = Y^{t_0}) = \tag{4.7}$$
$$\{y : y \leqslant (Y^{t_0}, Y^{t_0+1}, ..., Y^t) \cdot \lambda,\ x \geqslant (X^{t_0}, X^{t_0+1}, ..., X^t) \cdot \lambda,\ \lambda \geqslant 0\}.$$

The corresponding production set will be the set $\{(x, y) : y \leqslant (Y^{t_0}, Y^{t_0+1}, ..., Y^t) \cdot \lambda,$ $x \geqslant (X^{t_0}, X^{t_0+1}, ..., X^t) \cdot \lambda,\ \lambda \geqslant 0\}$. Therefore, the linear problem that defines the distance function relative to the sequential frontier becomes

$$\begin{gathered}\max_{\eta,\lambda\geq 0}\eta \\ \text{subject to} \\ -y\eta + (Y^{t_0}, Y^{t_0+1}, ..., Y^t) \cdot \lambda \geqslant 0 \\ x - (X^{t_0}, X^{t_0+1}, ..., X^t) \cdot \lambda \geqslant 0.\end{gathered}$$

The outcome of the latter problem can be used in (4.4) and (4.5) to compute the sequential Malmquist index and its decomposition. The component $TECH$ thus obtained shows pure technical progress and never indicates regress. All deteriorations in performance are attributed to the efficiency change component.[6] Since sequential

[6]Notice, since the construction of the conditional (or sequential) output set at time t uses information on all time periods within the interval $[t_0, t]$, the indices computed starting from different periods will have different sized reference sets. In practice, however, as $t - t_0$ increases, the distinction vanishes.

DEA uses past information to construct the frontier, the results of the sequential method are less sensitive to data attrition than the results of the contemporaneous method.

4.2.4 Synthesis of the two approaches

Consider the following example.

Example 4.1 *There are two countries A and B using the same quantity of input in year t and year t + 1 (that is* $x_A^t = x_A^{t+1} = x_B^t = x_B^{t+1}$*) to produce a single output y. Country A produces 1 unit of output in each year t and t + 1* ($y_A^t = y_A^{t+1} = 1$*), while country B reduces its production from 3 units of output in year t to 2 units in year t + 1* ($y_B^t = 3$, $y_B^{t+1} = 2$*). Since country B produces more output given the amount of input, it determines the production frontier in both years.*

Now let us compute the Malmquist productivity indices for both countries in this example. Country A's production has not changed between two years, therefore, $M_A(t, t+1) = 1$ for both contemporaneous and sequential methods. However, despite the fact that the two Malmquist indices are equal, their decompositions to technical change and efficiency change are different. It appears that in the case of contemporaneous frontiers efficiency improvement is offset by a negative shift of the technology: $M^A(t, t+1) = 1 = TECH \times EFFCH = \frac{2}{3} \times \frac{3}{2}$, while in the case of sequential frontiers both components show no change: $M^A(t, t+1) = 1 = 1 \times 1$. For country B the story is similar. $M^A(t, t+1) = \frac{2}{3}$ for both methods. However, depending on the choice of the reference frontier - contemporaneous or sequential - it is decomposed as $\frac{2}{3} \times 1$ and $1 \times \frac{2}{3}$, which means that the productivity change is interpreted as a pure technical change in the case of the contemporaneous Malmquist index and as a pure efficiency change in the case of the sequential index.

Note that in our example the contribution of a shift of the contemporaneous frontier relative to the sequential frontier is $\frac{2}{3}$. In the case of contemporaneous frontiers this shift is allocated to the technical change component, while in the case of sequential frontiers it belongs to the efficiency change. And this is exactly what causes the differences between the two decompositions. If we separate this factor and consider the combination of three shifts, shift of the sequential frontier, shift of the contemporaneous frontier relative to the sequential frontier and shift of an observation relative the contemporaneous frontier, we obtain that $M^A(t, t+1) =$

$1 = 1 \times \frac{2}{3} \times \frac{3}{2}$ and $M^B(t, t+1) = \frac{2}{3} = 1 \times \frac{2}{3} \times 1$. Or, more generally $M = TECH_S \times \frac{2}{3} \times EFFCH_C$.

Consequently, the formulae for the decompositions of the Malmquist indices can be rewritten as

$$M_C = TECH_S \times \frac{TECH_C}{TECH_S} \times EFFCH_C \tag{4.8}$$

$$M_S = TECH_S \times \frac{EFFCH_S}{EFFCH_C} \times EFFCH_C \tag{4.9}$$

Here and below the subscript C refers to the contemporaneous frontier and the subscript S refers to the sequential frontier. Consequently, the contemporaneous efficiency will be denoted as θ_C, while for the sequential measure we will use the notation θ_S.

It has been explained that the first factor in either of the above decompositions - the technical change component computed using sequential frontiers - reflects pure improvements of the technology ('technical progress'). The third one - contemporaneous efficiency change - shows changes of the gap between the leaders and the followers ('catch up').

The second factors in (4.8) and (4.9) correspond to shifts of the contemporaneous frontier relative to the sequential frontier or, in other words, changes of the position of the contemporaneous best practice relative to the best practice frontier ever achieved so far. This component measures productivity change attributable to the 'business cycle' via capacity utilization and labor hoarding.

Note that the two decompositions (4.8) and (4.9) are equivalent if and only if $\frac{TECH_C}{TECH_S} = \frac{EFFCH_S}{EFFCH_C}$, that is, when the measure of shifts of the contemporaneous frontier relative to the sequential frontier in (4.8) is the same as that in (4.9).

Obviously this component drops out if contemporaneous productivity sets are expanding (or at lest not shrinking) 'everywhere' over time. Then contemporaneous frontiers coincide with the corresponding sequential ones, and both decompositions (4.8) and (4.9) lead to the same result.

Proposition 4.1 *If for all* t, $t = t_0, t_0+1, ..., T$, *the output sets satisfy* $\{(x,y) : y \leqslant Y^t\lambda, x \geqslant X^t\lambda, \lambda \geqslant 0\} \subseteq \{(x,y) : y \leqslant Y^{t+1}\lambda, x \geqslant X^{t+1}\lambda, \lambda \geqslant 0\}$, *then* $M_C = M_S = TECH_S \times EFFCH_C$ *for all* t, $t = t_0, t_0+1, ..., T$.

Hicks-neutrality[7] provides another (sufficient) condition for the two indices to coincide.

Proposition 4.2 *If the technology exhibits CRS and there exists an output set* $\widehat{P}(x)$ *that all output sets* $P^t(x)$, $t = t_0, t_0+1, ..., T$, *satisfy the condition*

$$P^t(x) = A_t \widehat{P}(x), \tag{4.10}$$

in which $A_t \in \Re_+$, *then* $\frac{TECH_C}{TECH_S} = \frac{EFFCH_S}{EFFCH_C}$ *and* $M_C = M_S$.

Let me now turn to the empirical part. The next two sections present the data and the results.

4.3 Data

The data used in this study come from the International Sectoral Data Base (ISDB) constructed by the OECD statistical division. The ISDB contains a number of data series on sectoral outputs and primary factor inputs in 14 OECD countries (G7 and seven other countries, namely Australia, Netherlands, Belgium, Denmark, Norway, Sweden and Finland). The data are reported with annual frequency. The longest time series in ISDB cover the period between 1960 and 1995. However, for some countries the observation of the first ten years as well as the last few years are missing, which prompted the truncation of the time period in the analysis to 1970-1990. Moreover, three countries (Australia, The Netherlands and Norway) had to be dropped, because the data were missing for some years and industries.

The study covers the following manufacturing sectors:

- FOD - Food, beverages, tobacco;
- TEX - Textiles, wearing apparel and leather industries;
- CHE - Chemicals and chemical petroleum, coal, rubber and plastic products;
- MNM - Non-metallic mineral products except products of petroleum and coal;
- BMI - Basic metal industries;

[7] Note, we assume CRS. Condition (4.10) is the condition of Joint Hicks Neutrality for the CRS technology discussed in Färe and Grosskopf (1996).

- MEQ - Fabricated metal products, machinery and transport equipment.

Three categories of data are required: data on output, capital, and labor. Industrial value added[8] (series 'GDPD' in the ISDB classification) is taken as output, gross capital stock ('KTVD') as capital and total employment ('ET') as labor. Industrial value added presented in the ISDB is computed on the base of national accounts. Gross capital stock is estimated by means of a perpetual inventory model. Both data on output and capital are given in constant prices and in US dollars corresponding to 1990 purchasing power parities.

4.4 Empirical results

In this section I present the empirical findings. Subsection 4.4.1 summarizes the results on sequential and contemporaneous indices and their decompositions. In 4.4.2 I study the evolution of average efficiency in different sectors and identify the leaders in productivity. I also provide some evidence on the issue of convergence in TFP on sectoral level.

4.4.1 Analysis of the results on Malmquist indices

First, I compare the Malmquist indices based on the two alternative DEA models. Figure 4.1 shows the evolution of average Malmquist indices in each industry.[9] Here and below the average is computed by means of weighted geometric means. The solid line corresponds to contemporaneous frontiers and the dotted line to sequential frontiers. We can see that the two lines almost coincide, which indicates that the two measures of TFP growth produce very close results. However, this is not the case for the components in the decompositions of the Malmquist indices. Figure 4.2 demonstrates that the technical change components associated with the alternative approaches behave differently. The contemporaneous measure $TECH_C$ shows much more volatility than the sequential one. This is because it classifies each change in productivity of countries that belong to the frontier as technical change. On

[8]The ISDB gives value added in market prices. The rate of indirect taxes is also included in the ISDB, but it is missing in many cases. In this work no adjustment for indirect taxes has been introduced.

[9]The graph for 'MAN' presents the results for total manufacturing. 'SUM' is used for the sum of six studied industries.

the contrary, $TECH_S$ registers only those changes that lead to the expansion of the production possibility set. For example the two oil crises of 1973 and 1979, which caused overall fall in productivity, appear as declines in $TECH_C$, but have no impact on $TECH_S$.

Table 4.1 summarizes the correlations between the Malmquist indices and between their components. The first column shows correlations between the Malmquist indices: all numbers there are above 0.97. The next two columns correspond to efficiency change and technical change and give much smaller values than those in the first column. We observe that although the correlation between Malmquist indices is very high, the components show much less correlation. Thus, there is little discrepancy between the two Malmquist indices while there are significant differences in their decompositions. This agrees with our earlier finding from the analysis of Figures 4.1 and 4.2: the indices of TFP growth are very close, however their decompositions provide different interpretations to the sources of productivity growth. This is because contemporaneous indices, $TECH_C$, classify each change in productivity of the frontier countries as technical change. Thus they cover both forward and backward shifts of the frontier. In contrast, sequential indices, $TECH_S$, register only those changes that lead to the expansion of the production possibility set. The other changes are attributed to catch-up and reflected in $EFFCH_S$.

Table 4.1: Summary of correlations between the alternative Malmquist indices and their components

industry	$cor(M_C, M_S)$	$cor(EFFCH_C, EFFCH_S)$	$cor(TECH_C, TECH_S)$
FOD	0.971	0.807	0.756
TEX	0.979	0.899	0.712
CHE	0.989	0.444	0.336
MNM	0.991	0.657	0.556
BMI	0.985	0.727	0.470
MEQ	0.984	0.561	0.522

Figure 4.1: Evolution of the contemporaneous and sequential Malmquist indices

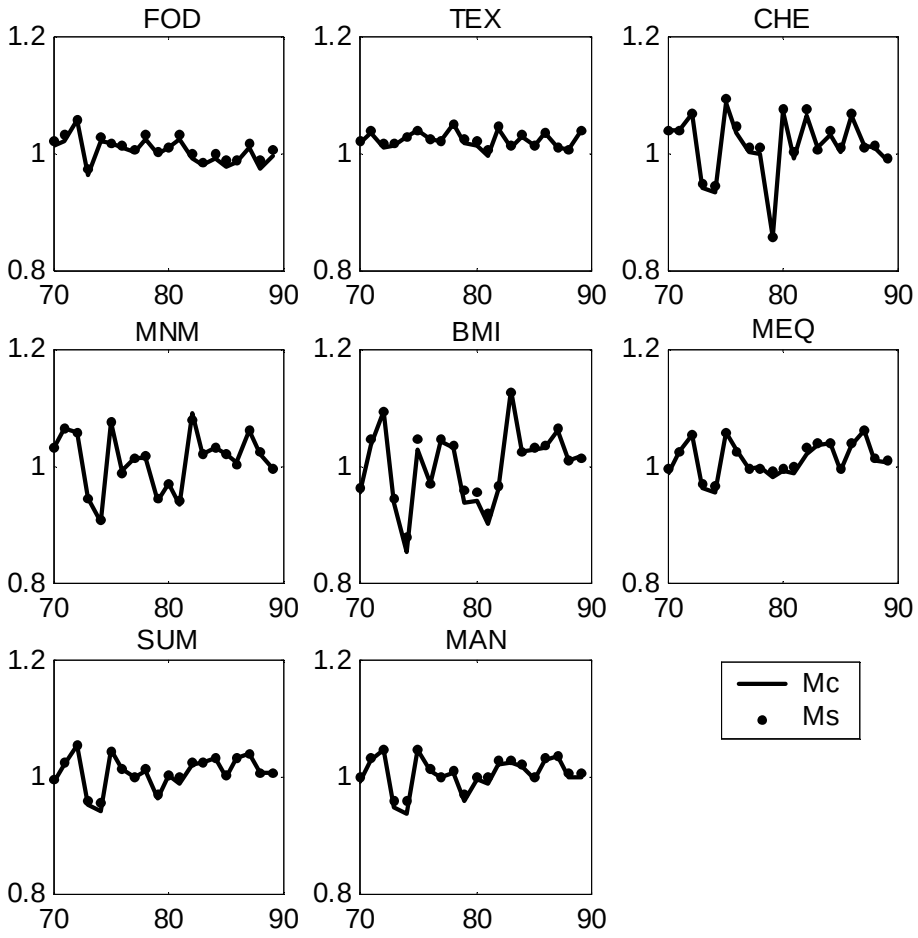
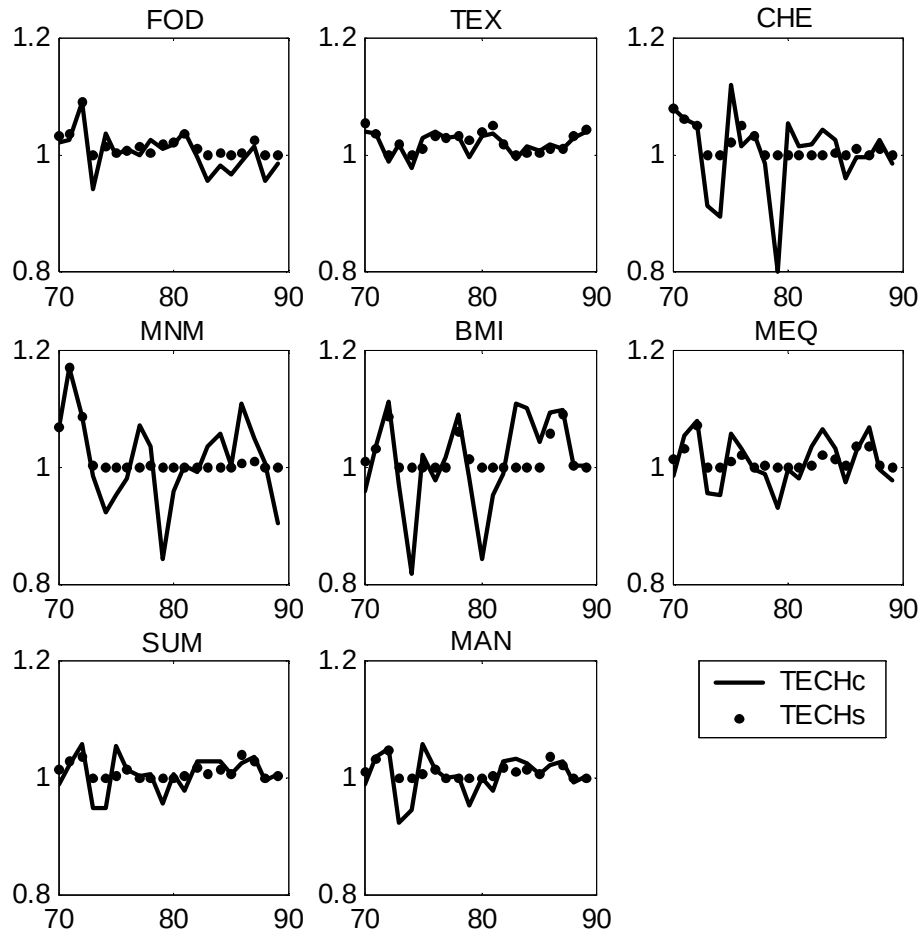


Figure 4.2: Indices of technical changes ($TECH$) measured as shifts of the contemporaneous and the sequential frontiers correspondingly



In Table 4.2 we compare the Malmquist indices and the corresponding technical change and efficiency change components for two subperiods: 1970-80 and 1980-90. According to both indices, textile industry experienced the highest TFP growth over the whole period. Although the technical change component was especially high in the first subperiod, four out of six industries showed a better performance in the eighties. This later subperiod is characterized by a somewhat higher catch up than the first subperiod. Notice also that in half of the cases we obtain $TECH_C$ less than one, indicating the average decline of productivity of the leaders in the corresponding industries.

Table 4.2: Comparison of the Malmquist indices and their components for subperiods 1970-80 and 1980-90

1970-1980

industry	M_C	M_S	$TECH_C$	$TECH_S$	$EFFCH_C$	$EFFCH_S$
FOD	1.013	1.018	1.015	1.021	0.998	0.997
TEX	1.026	1.029	1.019	1.024	1.007	1.005
CHE	0.993	0.999	0.984	1.027	1.010	0.973
MNM	1.000	1.001	1.004	1.029	0.996	0.972
BMI	0.984	0.995	0.992	1.020	0.992	0.975
MEQ	1.004	1.007	0.999	1.014	1.005	0.992

1980-1990

industry	M_C	M_S	$TECH_C$	$TECH_S$	$EFFCH_C$	$EFFCH_S$
FOD	0.995	1.002	0.991	1.011	1.004	0.992
TEX	1.020	1.023	1.018	1.022	1.002	1.001
CHE	1.008	1.012	0.991	1.002	1.018	1.010
MNM	1.007	1.008	0.993	1.001	1.015	1.006
BMI	0.999	1.006	1.013	1.015	0.986	0.991
MEQ	1.016	1.021	1.007	1.011	1.009	1.009

Table 4.3 presents the numerical results of the decomposition of TFP growth indices outlined in (4.8) and (4.9) over the period of 20 years. The average numbers are computed by means of weighted geometric means over the period. The last column in the table is given for the reader's convenience, to facilitate the comparison between the three-term decomposition of the Malmquist indices and their two-term decomposition (4.5). The highest TFP growth was observed in textile, machinery and chemical industries, and the lowest in basic metal products. Most of TFP growth is attributed to technical progress, the contribution was about 1.5-2% in all industries. The contribution of catching up was modest in most sectors, and even negative in the case of basic metal industry. Only in chemicals have we found a strong effect of catching up (1.3%). Therefore, for the average of OECD countries, the productivity gains in manufacturing are due to technical progress. This result obtained for separate manufacturing industries agrees with the earlier finding by Maudos et al. (2000). The latter applied Malmquist indices to analyze aggregate TFP growth in OECD countries over the period 1975-1990 and concluded that most of the productivity gains in OECD countries are attributed to technical progress. The contribution of the business cycle component appeared to be negative in most cases. The factor $\frac{TECH_C}{TECH_S}$ was always less than $\frac{EFFCH_S}{EFFCH_C}$, which implies that M_S was above M_C.

Table 4.3: Decomposition of the Malmquist indices

industry	M_C	$EFFCH_C$	$\frac{TECH_C}{TECH_S}$	$TECH_S$	$TECH_C$
FOD	1.003	1.002	0.986	1.015	1.001
TEX	1.023	1.003	0.997	1.023	1.020
CHE	1.010	1.012	0.985	1.013	0.998
MNM	1.007	1.000	0.992	1.015	1.007
BMI	0.995	0.991	0.987	1.017	1.004
MEQ	1.012	1.005	0.995	1.013	1.008
industry	M_S	$EFFCH_C$	$\frac{EFFCH_S}{EFFCH_C}$	$TECH_S$	$EFFCH_S$
FOD	1.009	1.002	0.993	1.015	0.994
TEX	1.026	1.003	1.000	1.023	1.003
CHE	1.015	1.012	0.990	1.013	1.002
MNM	1.008	1.000	0.993	1.015	0.993
BMI	1.002	0.991	0.994	1.017	0.985
MEQ	1.016	1.005	0.998	1.013	1.003

As explained in section 4.2, changes in the position of the current productivity leaders relative to the sequential frontier are not necessarily changes in technology. More likely they are attributed to the cyclical processes in the economies. The corresponding component has been dubbed as 'business cycle' to emphasize its cyclical nature. Separating effects of technical changes from cyclical behavior is desirable for the correct interpretation of productivity changes, as well as for the correct measuring of technical progress.

Cycles are closely related to variations in capacity utilization, and so does our 'business cycle' (BC) component. The contemporaneous frontier shifts inward when the utilization of capacity in the best-practice countries decreases, and moves back, when it restores. I recognize, however, that the effect of capacity utilization on TFP is much more complex. In particular, changes in capacity utilization contribute to the efficiency change component as well.[10]

4.4.2 Evolution of efficiency

In this section I apply the DEA model considered above to analyze the evolution of efficiency in the selected sectors. Table 4.4 summarizes the results for the average efficiency θ_C in each sector for four periods: 1970-1975, 1976-1980, 1981-1985 and 1986-1990. From these results we can identify technological leaders. They are listed in Table 4.5 below. That table shows that in most cases the leaders keep their leading position over the whole 20-year period.

[10] Recently De Borger and Kerstens (2000) suggested a way of incorporating of capacity utilization variations in the Malmquist index, by separating the variation in capacity utilization from the efficiency change component.

Table 4.4: Contemporaneous efficiency[11]

	Bel	*Can*	*Den*	*Fin*	*Fra*	*WG*	*Ita*	*Jap*	*Swe*	*GB*	*US*
FOD											
1970-75	0.738	0.969	0.364	0.476	0.783	0.785	0.726	1.000	0.713	0.652	1.000
1976-80	0.782	0.953	0.404	0.436	0.806	0.794	0.768	1.000	0.646	0.638	1.000
1981-85	0.821	0.856	0.442	0.447	0.731	0.777	0.773	1.000	0.664	0.663	1.000
1986-90	0.901	0.992	0.510	0.481	0.735	0.858	0.838	1.000	0.744	0.778	1.000
TEX											
1970-75	0.654	0.963	0.728	0.595	1.000	0.807	0.824	0.503	0.964	0.934	0.777
1976-80	0.722	1.000	0.838	0.592	0.993	0.885	0.944	0.474	0.846	0.786	0.895
1981-85	0.782	1.000	0.917	0.651	1.000	0.803	0.912	0.512	0.725	0.801	0.908
1986-90	0.868	1.000	0.759	0.658	0.999	0.862	0.970	0.443	0.762	0.793	1.000
CHE											
1970-75	0.192	0.404	0.490	0.397	0.643	0.916	0.290	1.000	0.711	0.712	0.722
1976-80	0.313	0.451	0.627	0.421	0.764	0.997	0.465	1.000	0.713	0.800	0.684
1981-85	0.593	0.485	0.663	0.487	0.866	1.000	0.627	1.000	0.741	0.736	0.766
1986-90	0.755	0.545	0.629	0.541	0.880	1.000	0.825	1.000	0.751	0.836	0.925
MNM											
1970-75	0.628	1.000	0.744	0.609	0.778	0.781	0.638	0.792	0.753	1.000	0.818
1976-80	0.698	0.985	0.719	0.602	0.887	0.876	0.837	0.657	0.710	1.000	0.809
1981-85	0.939	0.943	0.713	0.748	1.000	0.950	0.836	0.788	0.828	1.000	0.817
1986-90	0.964	0.995	0.576	0.702	1.000	0.884	0.806	0.720	0.780	1.000	0.842
BMI											
1970-75	0.507	0.574	0.364	0.339	0.487	0.750	0.589	0.884	0.336	1.000	1.000
1976-80	0.652	0.573	0.309	0.385	0.516	0.885	0.526	0.970	0.343	0.978	0.922
1981-85	0.788	0.554	0.396	0.494	0.531	1.000	0.643	0.957	0.420	1.000	0.846
1986-90	0.786	0.499	0.390	0.447	0.476	1.000	0.571	0.788	0.393	1.000	0.657
MEQ											
1970-75	0.865	0.974	0.775	0.544	0.860	0.945	0.597	0.606	0.732	0.914	1.000
1976-80	0.936	1.000	0.742	0.552	0.916	0.976	0.707	0.607	0.676	0.750	1.000
1981-85	0.998	0.990	0.780	0.645	0.907	0.938	0.759	0.795	0.758	0.687	1.000
1986-90	0.932	1.000	0.658	0.746	0.909	0.920	0.808	0.873	0.742	0.747	1.000

[11] In Table 4.4 'Bel'=Belgium, 'Can'=Canada, 'Den'=Denmark, 'Fin'=Finland, 'Fra'=France, 'WG'=West Germany, 'Ita'=Italy, 'Jap'=Japan, 'Swe'=Sweden, 'GB'=Great Britain and 'US'=the United States.

Table 4.5: The leaders

industry	*the leading countries*
FOD	US, Japan, Canada*[12]
TEX	Canada, France, US*
CHE	Japan, West Germany*
MNM	GB, Canada, France*
BMI	GB, West Germany*, US*
MEQ	US, Canada, Belgium*

Table 4.6 shows the average efficiency level of the industries for both methods. The second column gives lower numbers indicating that the contemporaneous frontier was sometimes shifting back in each industry. The gap between the two efficiency measures is very small for the textile industry (less than 1 %), but rather high in chemicals, basic metal products, and in non-metallic mineral products (about 10%), suggesting more backward shifts of the contemporaneous frontier in the latter three industries comparing with the others.

Table 4.6: Average contemporaneous efficiency

industry	*contemporaneous*	*sequential*
FOD	0.888	0.851
TEX	0.843	0.834
CHE	0.820	0.707
MNM	0.826	0.737
BMI	0.831	0.733
MEQ	0.898	0.865

As noted in section 4.2, contemporaneous efficiency provides us with a natural framework for studying convergence. The issue of convergence in TFP goes back to Dowrick and Nguyen (1989), who first pointed out that not only aggregate labor productivity levels of industrialized countries converge over time, but also their TFP levels. (See also Wolff (1993) and Baumol et al. (1994), Carree et al. (2000) for discussion of convergence on both aggregate and sectoral levels.)

[12] A star next to a country name indicates that the country was leading not over the whole 20-year period.

Convergence means that observations move towards each other in the course of time. The distance function reflects the distance between observations and the frontier. If there is convergence in TFP, then the mean efficiency[13] in the industry should approach to one with time, while the coefficient of variation should decline (the so-called σ-convergence, this terminology is due to Barro and Sala-i-Martin, 1991). Figure 4.3 shows the evolution of the average efficiency and the corresponding coefficient of variation in each industry. Strong convergence of TFP levels is observed in chemicals. There are also some indications of convergence in food industry and machinery. The last two graphs labelled by 'SUM' and 'MAN' present the results for the total of the six considered industries and for total manufacturing correspondingly. For this sample of eleven countries over the considered period signs of convergence of TFP are present on the aggregate level as well.

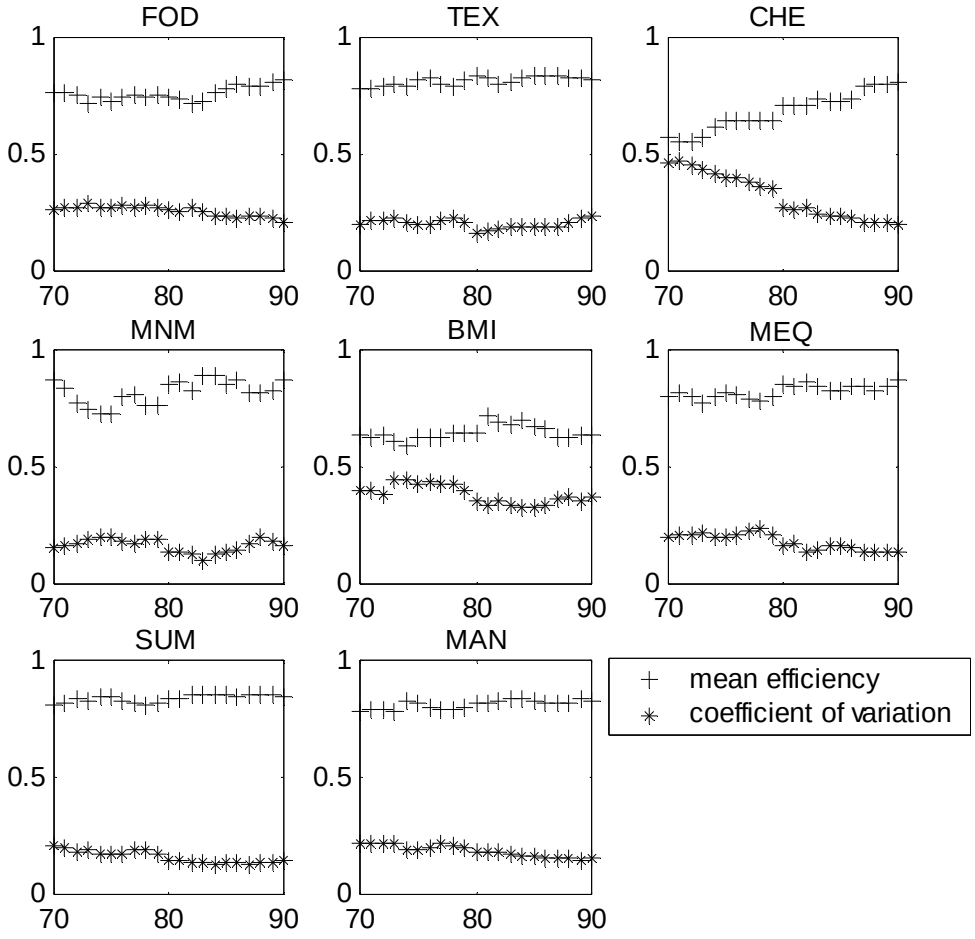
Another convergence hypothesis (β-convergence) asserts that a country with a lower initial TFP level should have a higher TFP growth. This implies that correlation between the initial efficiency and the subsequent indices of TFP growth should be negative. The correlation analysis has shown that this was the case in chemicals, food industry and total manufacturing. The correlations for these sectors are negative and significant at 95% level. Combining this result with the former, I conclude that these industries exhibit both σ-convergence and β-convergence.

4.5 Conclusion

In this chapter two approaches have been used to evaluate the TFP growth in manufacturing in eleven OECD countries, namely DEA with contemporaneous frontiers and DEA with sequential frontiers. It has been demonstrated that both methods produce highly correlated results for the total measure of TFP growth, but less correlated results for the decomposition into technical changes and efficiency changes. The sequential measure takes past information into account and reallocates temporary backwards shifts in the productivity of the best-practice countries to the efficiency change component, whilst the contemporaneous measure accounts for them as a technical regress. The former is more suitable for measuring technical changes in manufacturing.

[13] In this paragraph and in Figure 4.3 I use simple arithmetic means and the standard coefficient of variation, since they are commonly used in literature on convergence.

Figure 4.3: Evolution of the average efficiency and coefficient of variation in each industry



I suggest a decomposition of Malmquist indices, which links the two measures of TFP growth with each other. The new decomposition distinguishes three sources of TFP growth: technical progress, catching up and business cycle.

The empirical analysis has shown that most productivity increase in manufacturing in the OECD countries during the period 1970-1990 can be ascribed to technical progress. Five out of the six considered manufacturing sectors showed little or no catching up. Only in chemicals efficiency changes were substantial. I have found the strongest convergence of TFP levels for this sector. The contribution of the business cycle component of TFP growth appeared to be negative in most cases.

4.6 Appendix

Proof of Proposition 4.2

If condition (4.10) holds, then the distance functions computed relative to the contemporaneous frontiers satisfy the condition $D_o^t(x,y) = \widehat{D}_o(x,y)/A_t$. Therefore the formula for the contemporaneous Malmquist index can be rewritten as follows:

$$M_o(x^{t+1},y^{t+1},x^t,y^t) = \left[\left(\frac{D_o^t(x^{t+1},y^{t+1})}{D_o^t(x^t,y^t)}\right)\left(\frac{D_o^{t+1}(x^{t+1},y^{t+1})}{D_o^{t+1}(x^t,y^t)}\right)\right]^{1/2} =$$

$$= \left[\left(\frac{\widehat{D}_o(x^{t+1},y^{t+1})}{A_t}\frac{A_t}{\widehat{D}_o(x^t,y^t)}\right)\left(\frac{\widehat{D}_o(x^{t+1},y^{t+1})}{A_{t+1}}\frac{A_{t+1}}{\widehat{D}_o(x^t,y^t)}\right)\right]^{1/2}$$

$$= \frac{\widehat{D}_o(x^{t+1},y^{t+1})}{\widehat{D}_o(x^t,y^t)}.$$

Condition (4.10) implies that the sequential output sets satisfy $\overline{P}^t(x) = \max_{t_0\le s\le t} A_s \cdot \widehat{P}(x) = B_t \cdot \widehat{P}(x)$, where $B_t = \max_{t_0\le s\le t} A_s$, and consequently, distance functions based on sequential frontiers have to satisfy $\overline{D}_o^t(x,y) = \widehat{D}_o(x,y)/B_t$. Therefore, the formula for the sequential index can be reduced the same way as above, which completes the proof. ■

Part II

Incentive regulation and productivity performance

In the last few decades network industries such as electricity, gas and water supply have been undergoing major structural changes. In most European countries the utility sector, which has traditionally been a public monopoly, has been split vertically into separate segments: production, transportation over the network, and supply (or retail). While production and supply activities are considered to be competitive (at least potentially)[14], the transportation activity operated by regional monopolists requires government intervention. Two reasons: first, because network companies may extract monopoly rents from consumers; second, because shares of a network company can be held by supply and/or production companies and the network company may disadvantage competitors in production or supply. Therefore, a regulatory body is typically assigned to set a price control as to encourage the efficient use of the transportation system in the market context.

A transportation system includes a national transmission system operator managing the main grid (typical in electricity and gas) and many regional distribution companies transporting the commodity further to the final customers. Both the transmission and the distribution businesses are highly capital intensive. Facing uncertainty about future demand for transportation services, companies have to sink substantial investments before actual demand is realized. Once put in place, the grid serves customers for a long period of time (up to 50 years) and the capital costs are eventually recovered from customers. Customers often have to bear the risk of poor investment decisions by either paying an excessive price if the network has been goldplated (if there has been over-investment) or suffering from inadequate reliability of service if the installed capacity is insufficient.

In competitive industries with long-lasting investments, the most profitable company will be the one who predicted demand best and invested accordingly. Then the service is reliable, while per-unit cost is relatively low, and the company earns an adequate return on investment. In contrast, a company not utilizing its assets will not recover their fixed costs and will face a risk of bankruptcy. The market mechanism determines who gets rewarded or punished.

In regulated industries judgement is passed by a regulator. The regulator wants to prevent oversized projects, while giving companies incentives to operate efficiently and maintain a reliable service. His task in protecting the customers is to set a regime that would maintain a reasonable balance between the prices for transportation

[14]For a more detailed discussion on industries featuring both competitive and non-competitve components and seperation of these components, see OECD (2001).

services and the achieved level of reliability of supply. However, due to asymmetric information about demand, costs and technology, the regulator faces both moral hazard and adverse selection problems (Laffont and Tirole, 1993), which complicate his task.

This part of the thesis presents a model of regulation of regional network companies of the utility market. It starts with a review of the literature on regulation of natural monopolies, given in chapter 5, followed by the presentation of the model in chapter 6. The results of chapter 6 were first formulated in Mikkers and Shestalova (2001), in which we develop a framework that can be used by regulatory bodies to analyze the trade-off between prices and reliability of services (defined as the probability of interruption of the service) and demonstrate how the yardstick competition method can be augmented to factor in reliability of supply.

Chapter 5

Review of literature on regulation

In the early examples of regulation, franchised monopolies were typically subject to cost-of-service regulation. Under such a scheme revenues are set equal to costs (including a fair and reasonable rate of return) to eliminate the consumer welfare losses associated with monopolistic price distortions. However, as was noted in the sixties, this type of regulation does not motivate the firm to operate efficiently. On the contrary, since the offered rates of return are typically higher than the market cost of capital, firms have an incentive to overinvest in their assets - the so-called Averch-Johnson effect, first described by Averch and Johnson (1962) and then confirmed empirically by many other studies (see, e.g., Courville, 1974). As a consequence of this, public utilities tend to adhere to excessively high standards of reliability and uninterruptibility of service achieved by building costly and largely redundant networks (Kahn, 1995).

To mitigate inefficiency caused by a cost-of-service type of regulation the regulator can disallow 'not used and useful' investments to enter the rate base. According to Gilbert and Newbery (1988), this can overcome the firm's tendency to overinvest and can lead it to the choice of an efficient investment path in infinitely repeated regulatory interactions. Since the rate base (accumulated investments) determines the level of the firms' profit, its determination has been, as Kahn wrote, "by far the most hotly contested aspect of regulation, consuming by far the greatest amount of time of both commissions and courts". (Kahn, 1995, p. 36/I.) Due to asymmetric information about the cost and the effectiveness of investment, the regulator often cannot impose the desired level of investment on the firm.

Another problem with cost-of-service regulation was pointed out by Baumol and

Klevorick (1970), who showed that cost-of service regulation does not reward extraordinary entrepreneurial accomplishment. An inflexible application of cost-plus schemes might eliminate a financial reward for efficiency and innovation. Baumol and Klevorick (1970) suggested that introducing a regulatory lag could stimulate productivity improvement similarly to the mechanism of the Schumpeterian innovation process.

Littlechild (1983) formalized the regulatory lag and advised to the British Telecom regulator to introduce for the first time in regulatory practice the so-called price-cap scheme - a high powered incentive scheme in which the prices (in real terms) are fixed for a few years, giving the companies incentives to reduce the costs. Unfortunately, this did not completely solve the problem. Giulietti and Waddams-Price (2000) showed empirically that firms played strategic games anticipating price reviews. Giulietti and Waddams-Price came to the conclusion that firms regulated by price caps are not maximizing their profits within their regulatory constraints, because they act strategically and are more concerned with long-run issues of resetting the price cap.

As long as prices are directly linked to costs, incentives to operate efficiently remain weak. As pointed out by Joskow and Schmalensee (1986), price-cap regulation and cost-of-service regulation are similar in this respect. Bogetoft (2000) stressed that in a dynamic setting price-cap regulation does not guarantee cost efficiency, because the regulator has imperfect information about the cost functions of firms. Making judgements about the exogenous factors that influence the cost of a company, the regulator faces moral hazard and adverse selection problems. (See, e.g., Laffont and Tirole, 1993)

In competitive markets, price is a parameter to the seller - it is determined by market forces ('the invisible hand') and not subject to the individual seller's control, creating the pressure to improve productivity and quality of products/services. In contrast, the absence of competitive pressures leads in general to an increase of cost. (See Leibenstein, 1966.) Regulation (sometimes called 'the visible hand') can mimic the competitive forces. The regulator needs some benchmark other than the firm's present or past performance, against which to evaluate the firm's potential.

Shleifer (1985) proposed to use a yardstick competition regime, in which the regulator sets a price-cap for a firm based on average cost of the other companies and allows the firm to keep the difference between the cap and the realized cost. Since prices in this scheme are based on average cost of other companies, exogenous

shocks that affect the cost of the whole industry influence the price. The idea was developed further. In particular, Lyon (1991) applied it to evaluate disallowances of the recovery of construction costs. He suggested that hindsight reviews should be based on the lowest observable cost rather than on average other firms' costs. Furthermore, Bogetoft (1994, 1997) investigated the use of Data Envelopment Analysis in regulatory environments with technological uncertainty and showed that DEA-based reimbursement schemes ('DEA-based yardstick competition') may be optimal in this context.

Although there is a concern that sectors with an insufficient number of firms feature a risk of explicit or tacit collusion among participants (see, e.g., Tirole, 1988), even under such circumstances there are ways to make yardstick competition a success. For example, a collusion can be prevented by setting a reward for the firm that flags the coalition. (Bogetoft, 1995.)

The firms under yardstick regulation should face the same production opportunities and demand functions. In practice, yardstick competition can be implemented as long as the differences in circumstances are known. Still, some loss of incentive compared to a truly external competitive test arises due to the asymmetry of information. Each firm may argue that its costs are higher than that of its 'shadow firm', because of differences in circumstances. (Newbery, 1999.)

In the aforementioned yardstick competition schemes, the rewards depend on performance: if a firm outperforms the yardstick it earns a higher profit, otherwise it may incur losses. Yardstick competition thus mimics market forces and provides strong incentives to reduce cost and to improve efficiency. These cost-reducing incentives are especially strong in the short run, but unfortunately may have an adverse effect on investment in long-run objectives. In particular, a firm can delay an upgrade or the installation of new capacity, which may not affect today's performance, but result in the deterioration of performance in the future, affecting the quality of services (as seems to have happened in the case of the British Railways). To curb these undesirable effects price-cap regimes should be enhanced with some mechanism regulating quality.[1]

One aspect of quality, namely the reliability of service, is especially important in electricity regulation. Shortages occur due to the stochastic nature of demand:

[1]This holds also in static models without investment: if prices cannot be raised, but quality can be lowered (without the regulator noticing it), then quality will be lowered below the efficient level.

demand may sometimes exceed capacity, leading to an interruption of service. Two main issues arise. First, how much capacity to install, and second, how to ration the customers when an interruption occurs.

The existing literature presents results on possible price and rationing practices (see, e.g., Brown and Johnson, 1969, Jen and Tschirhart, 1979, Spulber, 1992), and emphasizes the importance of investment decisions and optimal capacity choice under uncertain demand. In particular, many authors highlight that excess capacity is not necessarily evidence of productive inefficiency, but may be an optimal response by a firm to uncertainty (Meyer, 1975, Nickel, 1978). As shown by Nickerson and Reynolds (1990), a non-negative response of optimal capacity to increased demand uncertainty may be demonstrated for a general class of consumer preferences.

On the other hand, changes in capacity, and thus reliability of services, also produce an impact on consumers demand. Coate and Panzar (1989) show that an increase of system reliability shifts the consumer demand curve outward and, therefore, that the optimal pricing rule should incorporate reliability. According to them, the optimal price should be equal to the marginal cost of providing another unit of electricity without degrading the quality (reliability) of service. The result of Coate and Panzar highlights the importance of the right mix between price and quality in the maximization of social welfare.

In the next chapter we analyze a similar issue and design a regulatory scheme which will resolve the trade-off between price and quality under information asymmetry.

Chapter 6

The model of yardstick competition of network utilities

6.1 Introduction

In this chapter[1] we develop a framework that can be used by regulatory bodies to analyze the trade-off between prices and reliability of services (defined as the probability of interruption of the service) and demonstrate how the traditional yardstick competition could be augmented to take reliability of supply into account. This can be achieved by implementing a simple performance based mechanism combining prices and penalties.

We consider a model of yardstick competition among distribution utilities operating in different regions but facing similar circumstances. Under demand uncertainty, given private information on parameters of the distribution of demand in the region, the companies choose the amount of capacity to install, as well as the amount of 'slack' (i.e. unnecessary spending that could have been avoided if more effort had been exercised). The installed capacity is related to the probability of failure and thus to the level of reliability. The regulator does not know the minimum installation cost and does not observe the amount of capacity installed directly, but has to induce the companies to operate efficiently and choose the socially-optimal capacity level.

We show that a yardstick competition scheme that does not penalize network failures, is suboptimal and leads to underinvestment. In contrast, the socially-

[1]The results presented in this chapter were first formulated in Mikkers and Shestalova (2001).

optimal outcome can be achieved by introducing penalties for undersupply equal to the value of the associated losses perceived by the customers[2]. Then the potentially external costs of inadequate supply are internalized by the companies and hence taken into account in making their investment decisions. Given that the regulator does not observe the firm's technology, the main problem is how to set prices that enable utilities to have sufficient expected revenue to cover both the efficient cost and the risk of shortfall, which will be reflected in fines that they occasionally have to pay. We solve this problem by introducing a yardstick competition regime augmented to incorporate the risk of network failure. Since we assume that all firms are able to achieve the same minimum installation cost, the proposed regulation scheme emerges as first best.

The chapter is organized as follows. We explain the model in section 6.2, which we solve in section 6.3. Section 6.4 provides a policy analysis, followed by a discussion of an application to yardstick competition in section 6.5. Section 6.6 concludes. All proofs are relegated to the appendix.

6.2 The model

In most of the European countries the electricity sector has been split into production, transmission, distribution and supply, each segment being operated by different firms. Here we focus on the distribution segment of the market, operated by regional network companies. Distribution companies maintain the regional network and deliver energy to the local customers.

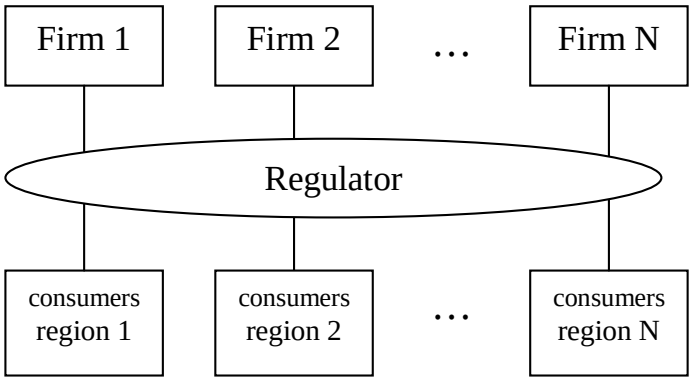
The structure of the distribution segment of the market is shown in Figure 6.1. Assume that there are N regions, in each of which there is a network company supplying this region with a particular service (such as distribution of electricity, gas or water). We denote the quantity delivered by y, $y \in R_+$. The provision of service y may be limited by the capacity of the regional distribution network, the installation and operation of which is under the firm's control.

Regional firms are monopolists and subject to regulation. Both the firms and the regulator face uncertainty about the realization of the consumers' demand for the service, however, the firms have private information about the range in which

[2]This result presumes risk-neutrality of both the firms and the customers.

demand will fall. The task of the regulator is to ensure efficient operation and to set prices maximizing the consumers' welfare under information asymmetry.

Figure 6.1: Distribution segment



The presentation of the model is organized as follows. We start with a description of the consumer preferences and the technology in sections 6.2.1 and 6.2.2, respectively. Then we touch upon the issue of the information asymmetry in section 6.2.3. Section 6.2.4 outlines the timing. Finally, section 6.2.5 refers to the problem of the regulator, which will be solved in section 6.3.

6.2.1 Consumer preferences

The utility that a consumer derives from consumption of distribution services depends on his demand for the distributed commodity itself. It reaches satiation as soon as the consumer's demand for the commodity at a fixed price is fulfilled. For example, in the case of electricity provision, the customer's utility from having an extra unit of available transportation capacity becomes zero as soon as he has already been provided with enough electricity. The satiation point is the point in which the demand for electricity (a commodity) is realized. It is subject to exogenous shocks, such as changes in fuel prices, structural changes in the region or weather conditions (adapting the use of electric heating or air conditioning). When supply is interrupted the consumers suffer, because they derive less utility than they would at the fulfilled demand.

To capture this, we assume that the consumers' preferences in region i, $i = 1, 2, ...N$ are represented by a utility function

$$U(m, y, Y_i, \kappa) = m + u(y, Y_i, \kappa)$$

where m is the numeraire commodity, so that income effect is absent. Consumers' benefits of distribution services have the form

$$u(y, Y_i, \kappa) = \kappa \min(y, Y_i)$$

where Y_i is a random variable with cumulative probability distribution $F(Y_i)$: $[a, b] \rightarrow [0, 1]$ and κ is a fixed parameter reflecting consumers' marginal willingness to pay for the service purchased from the regulated firm.[3]

Figure 6.2: Consumers' benefits

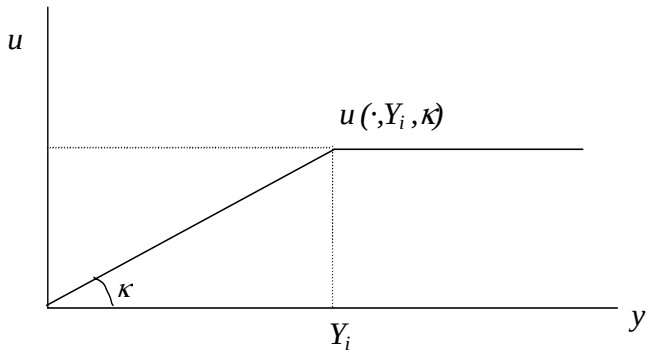
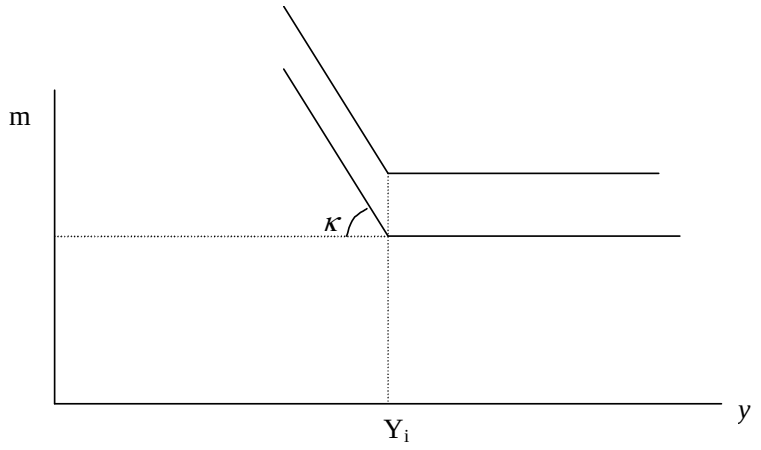


Figure 6.2 shows the consumers' benefits of distribution service. Function u increases at constant rate κ (see Remark 6.1 at the end of this section for a discussion

[3] A similar representation of the consumer preferences can be found in Spulber (1988, 1992). In Spulber's notations the consumer's utility function takes the form $\zeta + U(q, \theta, \omega)$, in which q is the amount of the good purchased from the regulated firm, θ is a parameter characterizing a type of consumer, ω is the state of world and ζ is a numeraire commodity. Spulber's $(\zeta, q, \theta, \omega)$ corresponds to our (m, y, κ, Y_i). The functional forms are different. In particular we assume satiation at Y_i and constant marginal willingness to pay for the service until the satiation is reached.

Figure 6.3: Indifference curves



of this assumption) until the satiation level Y_i has been reached and stays the same if the consumer is supplied with more than Y_i of this service. Consequently, demand for distribution services is perfectly inelastic and equal to Y_i as long as the price is below κ. For short we will call Y_i 'demand'.

If demand is fulfilled, then $u(y, Y_i, \kappa)$ equals $u_{\max}(Y_i, \kappa) = \kappa Y_i$. The expression for u can be rearranged as

$$u(y, Y_i, \kappa) = \kappa Y_i + \kappa \min(y - Y_i, 0) = u_{\max}(Y_i, \kappa) - \kappa(Y_i - y)_+$$

in which the second terms reflects consumer's disutility of nondelivered $(Y_i - y)_+ = \max(0, Y_i - y)$ units of service.

The consumer surplus, S_i, is the difference between the benefits of delivered services and the payment for it. Thus, $S_i = u(y, Y_i, \kappa) - R_i$, in which R_i is the revenue paid to the firm.

Remark 6.1 *The assumption of a constant marginal willingness to pay introduced above is driven by practical concerns. Function u considered above represents the utilities of the whole region and is the sum of individual utilities, which generally speaking may not be characterized by a constant marginal willingness to pay for the service. When the system fails, some consumers are rationed, while others are not,*

and (even in the case of identical customers) the losses caused by a shortfall will vary depending on the allocation of the lost load over the population of the region. A practical solution to this problem would be to average differences out by taking κ to be constant for each type of customers and to reflect their average disutility of an undelivered unit. Our model deals with one type of customer, but in principle can be generalized for the case of different types of customers with different (but constant) κ's, for example, industrial and residential customers. (This would imply a model of the rationing behavior of monopolists. See more on the latter issue in, e.g., Jen and Tschirhart, 1979, Coate and Panzar, 1989, Ahn et al., 1992, Spulber, 1992.) In practice, the marginal willingness to pay to avoid an interruption of supply can be estimated on the base of surveys.[4]

6.2.2 Technology

Among empirical papers analyzing the economies of scale in the utility sector there are both those supporting the economy of scale[5] and those which do not. Examples of the latter are Huettner and Landon (1978) and Kittelsen (1993)[6] on electricity transmission and distribution, Cubbin and Tzanidakis (1998)[7], Saal and Parker (2001)[8] on water-supply industry, etc. Although in theory monopolies are often characterized by increasing returns to scale (IRS), it is not the presence of IRS per se, but rather subadditivity of costs and restrictions on free entry and exit that make the market not competitive, as explained e.g. in Baumol et al. (1982).

The monopolistic character of the network businesses is due to large specific investment that firms have to sink before serving the customers. The investment

[4]For example, a series of surveys on customers' interruption costs was conducted in Norway in 1989-1991. Industrial, commercial and agricultural sectors were asked for direct costs associated with the interruptions of the power supply for specified number of hours. Residentilal customers were asked for willingness to pay to avoid interruptions. Basing on this the Norwegian regulator NVE estimated the average specific interruption cost per kWh, which was later used in regulation. (See Langset, 2001.) See also Caves et al. (1990) for a review of the outage cost literature.

[5]See, for example, Pollitt (1995), Dismukes et al. (1998), etc.

[6]Kittelsen (1993) has done a research on the Norwegian electricity distribution companies and found that the estimate of the VRS production set is indistinguishable from the CRS estimate for most of sizes observed.

[7]Cubbin and Tzanidakis (1998) have found modest economies of scale, insignificantly different from CRS.

[8]Saal and Parker (2001) estimate the scale elasticity to be in the range between 0.83 and 0.88 depending on the model used.

involves the installation of a fixed and unmovable connection between producer and customer, a duplication of which is inefficient. After the capacity has been put in place, the costs are sunk, and the provision of transportation services is done virtually at no cost[9], until the existing capacity is exhausted. Then a new installation takes place. In the long run, the firm adjusts the amount of capacity in accordance with the long-term trends in consumers' demand.

Assume that starting from some size Q_0 ($Q_0 < a$, where a is the lower limit of the support of the demand distribution) the technology for installation of capacity for provision of good y exhibits constant returns to scale (CRS) and is given by $I = (c + \delta)Q$. Here I is the amount of a numeraire good which has to be spent to install and operate capacity Q, c is a technical parameter reflecting minimal cost, which we assume to be the same in each region, and δ is a (per-unit) slack parameter. Having installed Q, a distribution utility can actually transport $\min(Y, Q)$ units of y, where Y is the demand for service y.

We define reliability of supply, ρ, as the probability of meeting the demand, that is, $\rho = \Pr\{Y < Q\}$. Therefore, each level of capacity Q corresponds to a certain level of reliability of supply ρ.

Suppose that firms can attract capital at rate r (r is the opportunity cost of capital), which is exogenously given. The firm's profit π is expressed as $\pi = R - (1 + r)I$, where R is revenue.[10]

Following Bogetoft (2000), we assume that the firms are risk neutral, seeking to maximize a weighted sum of their expected profit and slack. The firm's expected utility is represented by

$$E[U^f(\delta, Q)] = E[\pi + \omega(\delta Q)] = E[R] - (1 + r)I + \omega(\delta Q)$$

where ω is a fixed parameter that describes the firm's value of slack relative to profit. We assume $\omega \in [0, 1]$, that is, profit is more valuable to the firm than slack (because

[9]Since we focus on sunk costs problem here, we assume that the installed capacity can be operated and maintained at no cost. This simplifies the problem and allows us to avoid a discussion of allocation between fixed and variable costs, which is a really important issue in rate-of-return regulation, but has less relevance in our case, since prices are based on efficient total costs no matter what the allocation between fixed and variable cost is. Operation and maintenance cost can be easily incorporated in the model. The analysis still holds after the inclusion of these costs.

[10]Notice that cost of investment includes normal return on capital, r. Therefore, according to the terminology applied by business literature, π represents 'value creation'.

slack can be consumed only 'on-the-job' - Bogetoft, 2000, p.13). Notice that the inclusion of the utility of having slack has the same meaning as the inclusion of disutility of efforts: it costs efforts to eliminate slack. As Hicks noted: "The best monopoly profit is a quiet life". (Hicks, 1935, p.8.)

Without loss of generality we let the firm's reservation utility be 0.

We will also assume that $(1+r)cb < \kappa E[Y]$, where b is the upper bound of the domain of function $F(\cdot)$ introduced on p.78, and c is the minimal installation cost per unit of capacity, i.e., the first-best per-unit cost. That is, the consumer willingness to pay for having enough capacity to meet expected demand exceeds the first-best cost of installation of maximal relevant capacity.

6.2.3 Information asymmetry

We assume that all Y_i are independent of each other and drawn from the same distribution, with the cumulative probability distribution function $F(Y_i) : [a,b] \rightarrow [0,1]$ defined by $F(Y_i) = \widetilde{F}(\frac{Y_i - a}{b-a})$ for some distribution function $\widetilde{F} : [0,1] \rightarrow [0,1]$. Parameters a and b are the same across firms. They are known to the firms, but not to the regulator. This specifies the common information set of the firms: $\{\widetilde{F}, a, b\}$.

Furthermore, we consider network companies operating under similar circumstances. Therefore, we assume the technical parameter c to be the same for all firms, however, unknown to the regulator who observes only total cost $I = (c+\delta)Q$.

An important implication of our simplifying assumption of identical minimal costs is that the proposed yardstick competition scheme emerges as first best. If the number of observations is sufficient, the present analysis could be extended to incorporate the environmental differences, for example along the lines proposed by Shleifer (1985) or Bogetoft (1997). An extension to the case of different types of firms may be at the expense of shifting to the second-best outcome as in the standard Laffont-Tirole problem incorporating unobservable variation in types as well as efforts (see Laffont and Tirole, 1993, for more discussion) and will not be considered here.

The economic literature already offers schemes that would lead to the "full surplus extraction" (e.g., McAfee et al., 1989, McAfee and Reny, 1992), however, there are complications with their implementation in practice. Our approach is to construct a scheme that would be simple and practical.

6.2.4 Timing

Figure 6.4 shows the timing of the game. It is a one-period model: during the period the firms install capacity, provide the service and receive a payment.

First, the regulator offers a contract to the firms. A contract specifies the payment to the firm as a function of variables which will be observable at the end of the period. The functional form depends on the type of regulation chosen by the regulator.

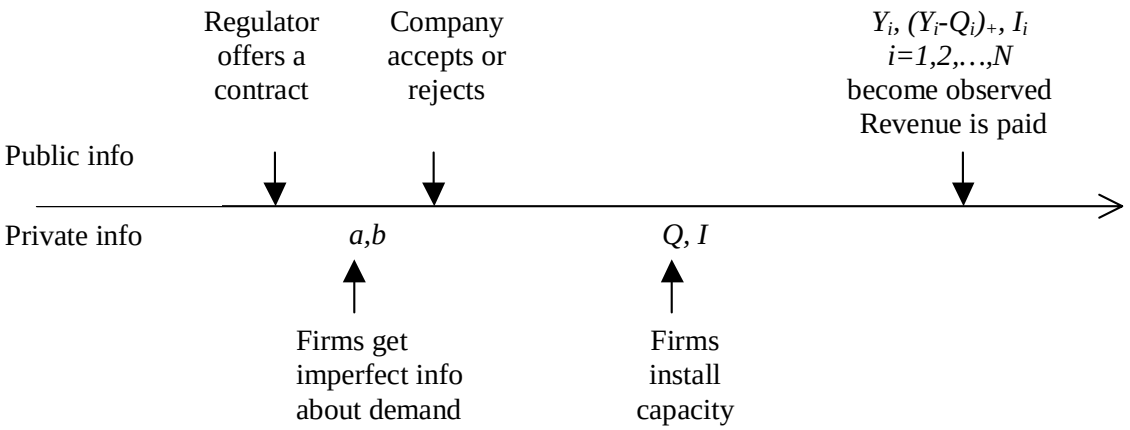
Second, the firms costlessly learn private information about the support of the distribution of demand, $[a, b]$, and decide to accept or reject the contract. After a contract is accepted a firm makes its investment decision and installs capacity. Notice that the companies need to make the investment decision and sink costs before the demand is realized. The investment is irreversible and will be completely depreciated at the end of the period.

Third, all Y_i, $(Y_i - Q_i)_+$ and I_i become observable[11]. The firm provides its service to the customers and receives revenue in return. Note that Q_i is not observable by the regulator.[12]

[11] Strictly speaking, what the regulator observes is the number of delivered units, that is $\min(Y_i, Q_i)$. Number of undelivered units $(Y_i - Q_i)_+$ is not directly observable, however it can be estimated by the regulator on the base of information on time of interruptions and typical load profiles of the affected customers over that time. Consequently Y_i can be obtained as $Y_i = \min(Y_i, Q_i) + (Y_i - Q_i)_+$.

[12] Obviously the installed capacity, Q_i, is not observable for those firms for which demand has not hit the capacity, $Y_i \leqslant Q_i$. But even if capacity is observable ex-post, there still exists informational asymmetry about what was the optimal capacity ex-ante, due to the private information about the demand distribution.

Figure 6.4: Timeline



6.2.5 Regulation

The task of the regulator is to design a contract, that would promote efficient operation and allocate the welfare gains to consumers. He seeks to prevent waste or misuse of resources and to assure cost control, in order to achieve efficient pricing of services and to eliminate welfare losses caused by price distortions. Thus, the regulator wants to maximize the consumer surplus (which is the same as to minimize the expected cost of making the firms to accept employment and disutility of losses) and to minimize the firms' incurred costs.

The contract specifies the payment to each firm i, $R_i = R(Y_i, (Y_i - Q_i)_+, I_i, Y_{-i}, (Y_{-i} - Q_{-i})_+, I_{-i})$, as a function of the observable data on firm i as well on all the other firms (for which we use notation $-i$) so as to maximize the consumers' expected surplus in the region, while ensuring a nonnegative expected profit to the regional company.

Since the demand distribution is conditional on a and b, the expected consumer surplus in the region is also conditional on them. The regulator does not observe the support of demand distribution, but has beliefs about a and b. We will show that it is possible to specify a regulation scheme that does not depend on the values of a and b.

We formulate the regulator's problem as follows

$$V(Q_i, \delta_i) = \max_{R, Q_i, \delta_i} E_{a,b} E_{Y|a,b}[S_i] \tag{6.1}$$

subject to

$$E_{Y|a,b}[R(Y_i, (Y_i - Q_i)_+, (c+\delta)Q_i, Y_{-i}, (Y_{-i} - Q_{-i})_+, I_{-i})] - \tag{IR}$$
$$-(1+r)(c+\delta)Q_i \geqslant 0$$

$$E_{Y|a,b}[R(Y_i, (Y_i - Q_i)_+, (c+\delta_i)Q_i, Y_{-i}, (Y_{-i} - Q_{-i})_+, I_{-i})] - \tag{IC}$$
$$-(1+r)(c+\delta_i)Q_i + \omega\delta_i Q_i \geqslant$$

$$\geqslant E_{Y|a,b}[R(Y_i, (Y_i - Q)_+, (c+\delta)Q, Y_{-i}, (Y_{-i} - Q_{-i})_+, I_{-i})] - \tag{6.2}$$
$$-(1+r)(c+\delta)Q + \omega\delta Q \quad \forall Q, \delta > 0.$$

Here $E_{Y|a,b} = E_{Y_1 \ldots Y_N|a,b}$ denotes the expectation conditionally to the parameters of the distribution of the random variable Y.

The first constraint is the individual rationality (IR) or, in other words, the participation constraint. It implies that the firm is willing to stay in business: it expects to at least earn its investment back and receive return on investment of at least r. The second, called incentive compatibility (IC), says that the maximum of the firm's utility is achieved when the firm chooses slack and the amount of capacity maximizing the consumers' welfare. The latter implies that the firm chooses the optimal reliability level.

Under the information asymmetry about a and b, the regulator can specify the contract (the expression for R). Suppose he offers firm i a contract, according to which the firm receives price p_i for each unit it delivers and has to pay to consumers compensation φ_i for each undelivered unit. Then revenue of the firm is given by

$$R(p_i, \varphi_i, \min(Y_i, Q_i), (Y_i - Q_i)_+) = p_i \min(Y_i, Q_i) - \varphi_i (Y_i - Q_i)_+.$$

We will show that for a certain specification of p_i and φ_i this contract will lead to the optimal outcome for any a and b.

6.3 Solving the problem

The problem described in section 6.1 will be solved in a number of stages.

First, we consider total welfare maximization and find the socially optimal amount of capacity to be installed, as well as the socially desirable level of reliability of supply. Then we analyze the problem of the consumers purchasing service y under the capacity constraint, and find the preferred level of capacity for any given p and φ. Finally, we consider the firm's problem and obtain a necessary condition on p and φ to achieve the socially desirable level of reliability of services. This condition will be used to express the participation constraint as a constraint on the minimum price.

6.3.1 Total welfare maximization

Let us consider a specific region. In this subsection we drop subscript i corresponding to the region to simplify the notation. For a and b, given capacity Q $(a \leqslant Q \leqslant b)$, the expected surplus of the consumers in this region is equal to

$$E[S] = E[\kappa Y - \kappa \min(y - Q)_+ - R] =$$
$$= \kappa E[Y] - \kappa \int\limits_{[Q,b]} (Y - Q) dF(Y) - E[R].$$

Here, $E = E_{Y|a,b}$.

The expected utility of the firm is

$$E[U^f] = E[R] - (1 + r)I + \omega\delta Q.$$

Therefore total welfare is determined by

$$E[S] + E[U^f] = \kappa E[Y] - \kappa \int\limits_{[Q,b]} (Y - Q) dF(Y) - (1 + r)(c + \delta)Q + \omega\delta Q. \quad (6.3)$$

The maximization of (6.3) with respect to (δ, Q) taking values from $R_+ \times [a, b]$ gives the following conditions on the maximum

$$\begin{aligned} \delta &= 0 \\ \kappa(1 - F(Q)) &= (1 + r)c, \end{aligned}$$

from which we obtain the socially optimal capacity level

$$Q^* = F^{-1}\left(1 - \frac{(1 + r)c}{\kappa}\right). \quad (6.4)$$

The conditions defining the maximum have the standard interpretation. First, it is never optimal to have slack δ positive. Second, the expected marginal benefits of having an extra unit of capacity are equal to the marginal cost of its installation.

Note that each level of capacity Q corresponds to a certain reliability of supply $\rho = \Pr\{Y < Q\} = F(Q)$. From (6.4) we can obtain that the socially optimal reliability of supply is determined as $\rho^* = F(Q^*) = 1 - \frac{(1+r)c}{\kappa} < 1$.

6.3.2 Consumer's surplus maximization under a capacity constraint

Let us now consider the optimal outcome from the point of view of consumers alone.

Assume that the region i purchases service from the regulated firm at price p_i and gets compensated for each undelivered unit by fixed amount $\varphi_i \geqslant 0$. The amount of y available for the purchase is restricted by the amount of capacity installed, Q_i.

Given p_i, φ_i and the available capacity Q_i $(a \leqslant Q_i \leqslant b)$ the expected consumer's surplus conditional on (a, b) is

$$\begin{aligned} E_{Y|a,b}[S_i] &= \kappa E[Y] - E_{Y_{-i}|a,b} \int\limits_{[a,b]} (p_i \min(Q_i, Y_i) + (\kappa - \varphi_i)(Y_i - Q_i)_+) dF(Y_i) = \\ &= \kappa E[Y] - E_{Y|a,b}[p_i Y_i] - E_{Y_{-i}|a,b} \int\limits_{[Q_i,b]} (\kappa - \varphi_i - p_i)(Y_i - Q_i) dF(Y_i). \end{aligned}$$

Here notation $-i$ stands for all the other regions $j = 1, ...N$, $j \neq i$.

If $p_i + \varphi_i < \kappa$ this function is maximized at $Q_i = b$. If $p_i + \varphi_i > \kappa$ the maximum is reached at $Q_i = a$. As shown in the previous section, these levels of capacity do not maximize total welfare.

In the special case in which $p_i + \varphi_i = \kappa$, the desirable Q_i is any number within the interval $[a, b]$. The consumer's compensation for nondelivery is exactly equal to the utility loss caused by it and the surplus is always expressed by $S_i = (\kappa - p_i)Y_i$, no matter what amount of capacity is available.

6.3.3 Problem of the firm

In this subsection we consider the solution of the regulated firm's profit maximization problem and combine it with the results of social welfare maximization. This will allow us to formulate the conditions on prices and fines required to achieve the social optimum.

If the regulator offers a firm a contract, according to which the firm receives price p for each unit it delivers and has to pay to consumers compensation φ for each undelivered unit, then the firm with capacity Q gets utility $U^f = p\min(Y,Q) - \varphi(Y-Q)_+ - (1+r)I + \omega\delta Q$.

Suppose the regulator delinks prices and fines from the own cost of the firm. That is, when designing the optimal contract, he specifies p and φ as functions of the observed performance of other firms, not the performance of the firm itself. Then for each firm i the correspondent p_i and φ_i are independent of the investment incurred by firm i, I_i. We will show this will motivate the firm to eliminate slack.[13]

Given the specification of p_i and φ_i, the firm i's expected utility is as follows

$$\begin{aligned} E_{Y|a,b}[U_i^f](\delta_i, Q_i) &= E_{Y|a,b}[p_i\min(Y_i,Q_i) - \varphi_i(Y_i-Q_i)_+] - \\ &\quad -(1+r)(c+\delta_i)Q_i + \omega\delta_i Q_i = \\ = E_{Y_{-i}|a,b}&\left[\int\limits_{[a,b]} p_i Y_i dF(Y_i) - \int\limits_{[Q,b]} (p_i+\varphi_i)(Y_i-Q_i))dF(Y_i)\right] - \\ &\quad -(1+r)cQ_i + (\omega - r - 1)\delta_i Q_i. \end{aligned}$$

If $p_i + \varphi_i$ is a constant, then FOCs with respect to (Q_i, δ_i) give us

$$\begin{aligned} \delta_i &= 0 \\ (p_i+\varphi_i)\int\limits_{[Q,b]} dF(Y_i) - (1+r)c &= 0, \end{aligned}$$

from which we obtain

$$1 - F(Q_i) = \frac{(1+r)c}{p_i+\varphi_i}, \tag{6.5}$$

and therefore, the amount of capacity that maximizes firm i's expected utility is

$$Q_i^f = F^{-1}\left(1 - \frac{(1+r)c}{p_i+\varphi_i}\right). \tag{6.6}$$

Notice that to derive the above formula we used that the sum $p_i + \varphi_i$ is a constant, and that p_i and φ_i are independent on the actions of company i. However, p_i and φ_i may depend on the realization of demand in region i.

[13] In contrast, in case of a rate-of-return regime the allowed revenue would be set equal to the incurred cost, i.e. $R = (1+r)I$, which would lead to $\pi = 0$ and $U^f = \omega\delta Q$, giving no incentive to eliminate slack.

Remark 6.2 *The firm's problem considered here has much in common with the so-called 'newsboy's problem' arising in inventory models, which can be formulated as follows: given that the total demand over the period is uncertain the dilemma is to order enough inventory, so that the full potential profit may be realized, but not too much, so as to avoid losses in excess. (See, e.g., Ravindran et al., 1987, p.353-356.) The solution of this problem is called a critical ratio policy, since the optimal amount of inventory depends on the ratio of the potential loss per unsold item over the sum of potential profit per item sold, loss per item of unmet demand and loss per unsold item. Our term* $\frac{(1+r)c}{p+\varphi}$ *can be interpreted along the same lines, if we rewrite the denominator as* $p + [\varphi - (1+r)c] + (1+r)c$.

Q^f defined above is an increasing function of both p and φ. That is, a higher price set by the regulator as well as larger fines for nondelivery will motivate the firm to install more capacity. If the regulator knows the firm's production function (i.e. parameter c) and the opportunity cost of capital r, he is able to enforce any level of reliability of supply by simply choosing an appropriate pair of p (p does not depend on the actions of the company) and φ, summing up to the corresponding number in accordance with (6.6). For example, if the regulator would like to achieve reliability of supply equal to ρ, he should solve $1-\rho = \frac{(1+r)c}{p+\varphi}$ and set $p+\varphi = \frac{(1+r)c}{1-\rho}$. In particular the following proposition holds.

Proposition 6.1 *The socially optimal level of capacity* Q^* *defined by (6.4), can be achieved by setting* p *and* φ *such that* $p + \varphi = \kappa$.

This result resembles the well-known notion of a Pigovian-tax (see Pigou, 1920). The correct amount of capacity can be ensured by penalizing undersupply by the value of lost load perceived by the consumers, so that the potentially external cost of inadequate supply is internalized by the firms.

No information other than κ will be necessary to enforce efficient operation. As long as prices are delinked from cost, the firm will have an incentive to eliminate slack and install the socially optimal amount of capacity. Notice that maintaining the condition $p + \varphi = \kappa$ will not only lead to maximization of total welfare, but will also protect the consumer from the risk of nondelivery, making sure that at any p_i the consumer's surplus is always $(\kappa - p_i)Y_i$.

6.3.4 Participation constraint

The split of κ between p and φ ($p + \varphi = \kappa$) is a sheer distributional issue. As p increases, so does the share of total surplus allocated to the firm. Since the regulator protects the interests of consumers and firms, he would like to minimize payments to the firm, while assuring that the firm earns a nonnegative expected profit. Given imperfect information the derivation of p becomes a complex task.

Notice that after slack has been eliminated, the firm's expected utility becomes equal to the expected profit. The firm will enter the contract as long as there exists a level of investment at which its expected profit is nonnegative. Therefore, the participation constraint of firm i is reduced to $E_{Y|a,b}\pi_i(0, Q_i^f) \geqslant 0$, from which we can derive a constraint on the minimum price.

Starting with

$$E_{Y|a,b}[\pi_i](0, Q_i^f) = \tag{6.7}$$

$$= E_{Y_{-i}|a,b}\left[\int\limits_{[a,b]} p_i Y_i dF(Y_i) - (\varphi_i + p_i)\int\limits_{[Q_i^f,b]} (Y_i - Q_i^f)dF(Y_i)\right] -$$

$$-(1+r)cQ_i^f \geqslant 0$$

and taking into account that in the welfare maximizing case $p_i + \varphi_i = \kappa$ and $Q_i^f = Q^*$, we obtain that

$$E_{Y|a,b}[\pi_i] = E_{Y|a,b}[p_i Y_i - \kappa(Y_i - Q^*)_+] - (1+r)cQ^* \geqslant 0$$

and consequently

$$E_{Y|a,b}[p_i Y_i] \geqslant (1+r)cQ^* + \kappa E_{Y|a,b}(Y_i - Q^*)_+. \tag{6.8}$$

Due to demand uncertainty the minimum $p_i Y_i$ is above the cost of capacity to compensate for the fines that the firm may occasionally have to pay.

6.4 Policy analysis

In the previous section we concluded that the firm's choice of capacity depends on the sum of prices and fines. In this section we discuss the impact of varying the level of the fine for non-delivery on the choice of capacity. The table presented at the end of this section summarizes the results of the discussion.

6.4.1 Case of φ=0 (no fines)

If φ=0, then the price is the only instrument available to the regulator to control the behavior of the firm.

If price is set at the lowest possible level, just to cover the cost, $p = (1 + r)c$, then according to (6.6) the firm response will be to install $Q^f = a$. This will lead to the lowest possible quality, since according to our assumptions the realization of Y will always be greater or equal to a.

At any $p < \kappa$ the capacity is less than socially optimal. Firms underinvest. By increasing the price up to κ the regulator can enforce the socially optimal capacity level and thus to maximize the total welfare. However, at $p = \kappa$ all surplus goes to the firm, leaving the consumer with zero net surplus from having this firm.

6.4.2 Case of $\varphi \geqslant \kappa$

Since $p > 0$, then $p + \varphi > \kappa$, it follows that $Q^f > Q^*$ and the firm will overinvest. Nevertheless, only by setting infinitely large fines the regulator ensures that the firm chooses to install the highest level of capacity (i.e. the capacity corresponding to the 100% reliability), since $Q^f < b$ for any finite φ. If the regulator has no information on κ and considers perfect reliability as being the ideal outcome for customers, he may set severe punishment for nondelivery and force the firms to install full capacity. Let us consider this special case as an illustration.

If $\varphi = \infty$, then $Q^f = b$ and $E[\pi] = E[pY - (1 + r)cb]$. The firm breaks even if

$$p = (1 + r)cb/E\,[Y] = p_{\min}. \tag{6.9}$$

Note that the regulator can observe the realized values of I and Y for all the firms, but does not have information on the technological parameter c and the upper boundary of the demand distribution b. Since $(1 + r)cb/E\,[Y] = (1 + r)I/E\,[Y] \geqslant (1 + r)E[I/Y]$, the regulator can estimate $p_{\min}$ by taking a sample average of the other's firm observations. That is, for any i we obtain $p_i = (1+r)\frac{1}{(N-1)}\sum_{j\neq i}\frac{I_j}{Y_j}$. The term corresponding to the observation i is excluded from the sum, because above we assumed that p_i does not depend on Y_i for any i.

We should note that regulatory schemes assigning very high penalties for inadequate performance may be unrealistic in that they may bankrupt the company, which in practice most regulators would prefer to avoid. In the next section we turn to a more realistic case.

6.4.3 Case of $0 < \varphi < \kappa$

The socially optimal level of Q can be ensured by setting $\varphi = \kappa - p$. This will lead to $Q = Q^*$, $\delta = 0$ and to the maximum level of total welfare. To pass all the benefits to the consumers the regulator should set the price just to satisfy the participation constraint. We reformulate condition (6.8) as $p_{\min} = (1+r)I/Y + \kappa E(Y - Q^f)_+/Y$. Under incomplete information, the regulator has to estimate this value on the basis of the other firms' observed data.

The necessary information includes data on demand, undelivered units and investment. As before, this can be done by taking sample statistics. We will elaborate on this in the next section.

Table 6.4.1. Summary of the policy analysis

	$\varphi = 0$	$0 < \varphi < \kappa$	$\varphi \geqslant \kappa$
Q	underinvestment if $p < \kappa$	optimal if $\varphi + p = \kappa$	overinvestment
δ	eliminated	eliminated	eliminated
$p_{\min}$	$(1+r)c$	$(1+r)I/Y + \kappa E(Y - Q^f)_+/Y$	$(1+r)cb/E[Y]$

6.5 Discussion of yardstick competition under uncertain demand

In this section we discuss the application of yardstick competition and propose a compensation scheme that will encourage the elimination of slack and the installation of the socially optimal capacity.

In practice the true technology level of the firm is often unknown to the regulator. Therefore, c has to be estimated. Suppose there were no demand uncertainty, then given the data on cost and output, the regulator could define a reasonable yardstick for the price of firm i as

$$p_i = (1+r)\widehat{c}_i, \tag{6.10}$$

in which $\widehat{c}_i$ can be estimated from observations on the performance of other firms

in the same industry.[14] In particular, $\widehat{c}_i$ can be defined either as an average (as proposed in Shleifer, 1985)

$$\widehat{c}_i = \frac{1}{N-1} \sum_{j \neq i} \frac{I_j}{Y_j} \tag{6.11}$$

or as the minimum (as in Bogetoft, 1994)

$$\widehat{c}_i = \min_{j \neq i} \left[\frac{I_j}{Y_j} \right]. \tag{6.12}$$

Now let us suppose that future demand is uncertain. Then the output, which represents what has actually been delivered, is conditional on capacity and equal to $\min(Q_j, Y_j)$. Incorporating this adjustment in the above formulae, we obtain

$$\widehat{c}_i = \frac{1}{N-1} \sum_{j \neq i} \frac{I_j}{\min(Q_j, Y_j)} \tag{6.13}$$

or

$$\widehat{c}_i = \min_{j \neq i} \left[\frac{I_j}{\min(Q_j, Y_j)} \right]. \tag{6.14}$$

While both schemes provide an estimate of the true cost parameter c, neither will be optimal when there is uncertainty about demand. For example, if the compensation scheme is based on (6.14), the firm that achieved the minimum investment per unit delivered, and therefore has installed the least excess capacity, earns a non-negative profit, while profits of all the other firms are negative. As a response to such a compensation scheme regulated firms will eventually decrease capacity, to maximize the utilization of their assets. This means that the firms will underinvest. In particular, the following proposition can be proved.

Proposition 6.2 *A regulatory scheme combining 'no-fine'-regime with (6.10), in which $\widehat{c}_i$' s are defined by either (6.13) or (6.14) leads to underinvestment and consequently to low reliability of supply. In particular, it supports $Q_i = a_i$ $(i = 1, 2, ..., N)$ as a Nash equilibrium.*

[14] Alternatively, the regulator can use an engineering cost-proxy model as for example is applied by the Federal Communications Commission (FCC), the regulator of the telecom industry in the US. However, this approach does not eliminate the need for data on other firms. Cost-proxy models should be calibrated on other firms' data. (See Gasmi et al., 1999.)

As discussed in sections 3.3 and 3.4, the optimal level of capacity is achieved when $p+\varphi = \kappa$ and $pY \geqslant (1+r)I + \kappa E(Y-Q^f)_+$. Therefore, the regulator should set prices in accordance with this formula. Following Shleifer (1985), we apply the idea of yardstick competition based on the average of other firms' observations and estimate the latter expression by taking the sample statistics.

Proposition 6.3 *The optimal compensation scheme can be specified as follows: for every i*

$$R_i = p_i \min(Y_i, Q_i) - \varphi_i (Y_i - Q_i)_+,$$

in which

$$p_i = \min\{\kappa, (1+r)\overline{I}_{-i}/Y_i + \kappa(\overline{(Y-Q)_+})_{-i}/Y_i\} \tag{6.15}$$

$$\varphi_i = \kappa - p_i. \tag{6.16}$$

Here 'bars' denote average of the other firms' observations, for example,

$$\overline{x}_{-i} = \frac{1}{N-1}\sum_{j\neq i} x_j. \tag{6.17}$$

Since firms are exposed to the risk of a (stochastic) shortfall, the individual rationality constraint is not met if a firm does not receive a mark-up to compensate for this risk. The scheme specified in Proposition 6.3 guarantees the firm revenue $R_i = (1+r)\overline{I}_{-i} + \kappa(\overline{(Y-Q)_+})_{-i} - \varphi(Y_i - Q_i)_+$, which means that a firm should receive a price exceeding the minimum cost of installing a unit of capacity. The mark-up $\kappa(\overline{(Y-Q)_+})_{-i}$ is based on the average number of units non-delivered by the other companies. It is needed to compensate for the risk of fines. A company faces a trade-off between increasing reliability and the cost of installing extra capacity.

Note that the regulator is able to achieve the first-best outcome without actually knowing a, b and c. He only knows that a, b and c are the same for all firms and that the Y_i are independently and identically distributed.

Remark 6.3 *Proposition 6.3 can be generalized for the case of different-size regions, under the assumption that $\frac{a}{b}$ is constant across regions. Then the demand distribution in each region i of size s_i is given by $F_i(Y_i) : [a_i, b_i] \to [0,1]$, $F_i(Y_i) = \widetilde{F}\left(\frac{Y_i - a_i}{b_i - a_i}\right)$, where a_i and b_i are expressed as $a_i = as_i$, $b_i = bs_i$. The only adjustment to the scheme will be the incorporation of weights into formula (6.17) so that $\overline{x}_{-i} = \frac{1}{N-1}\sum_{j\neq i}\frac{s_i}{s_j}x_j$.*

6.6 Conclusion

Focusing on the regulation of regional distribution utilities operating under similar circumstances and able to achieve the same minimum cost, we show how traditional yardstick competition can be augmented to resolve the trade-off between price and reliability of service. We propose a regulation scheme that incorporates the aspect of capacity choice under uncertainty. The scheme allows the regulator to enforce the desired level of investment, corresponding to the socially-optimal level of reliability of supply, and allocate the welfare gains to the customers.

We show that a scheme that does not penalize network failures, is suboptimal and leads to underinvestment. In contrast, the socially-optimal outcome can be achieved by introducing penalties for undersupply equal to the customer value of the associated losses. Then the potentially external costs of inadequate supply are internalized by the companies and hence taken into account in making their investment decisions.

In our model companies are exposed to the risk of a shortfall. In order that the individual rationality constraint under demand uncertainty is met, the regulator should therefore offer a firm an expected price which exceeds the minimum cost of installing a unit of capacity. We have suggested how this markup, which is associated with the risk of undersupply might be quantified by a regulator when there is asymmetric information.

6.7 Appendix

Proof of Proposition 6.2

According to our assumptions $p < \kappa$ (otherwise there is no sense to have this firm at all), therefore, $Q^f < Q^*$ always holds, implying underinvestment. In particular, if all other firms j, $j \neq i$, invest a_j, both above formulae (6.13) and (6.14) lead to $\widehat{c}_i = c$, and consequently to $p_i = (1+r)c_i$, which can be sustainable only at $Q_i = a_i$.

■

Proof of Proposition 6.3

It has already been shown that as long as p_i is independent of the actions of the company i and the fines are set as $\varphi_i = \kappa - p_i$ the company will eliminate slack and

install the welfare maximizing amount of capacity. What is left to be shown is that p_i satisfies the participation constraint. Combining (6.7) with (6.15) and (6.16) we obtain

$$E_{Y_1 \ldots Y_N|a,b}\left[\pi_i\right] = \tag{6.18}$$

$$= E_{Y_{-i}|a,b}\left[\int\limits_{[a,b]} p_i Y_i dF(Y_i) - \int\limits_{[Q_i^f,b]} (\varphi_i + p_i)(Y_i - Q^*)dF(Y_i)\right] - (1+r)cQ^* =$$

$$= E_{Y_{-i}|a,b}\left[\int\limits_{[a,b]} \left((1+r)\overline{I}_{-i}/Y_i + \kappa(\overline{(Y-Q)_+})_{-i}/Y_i\right) Y_i dF(Y_i)\right] -$$

$$-\kappa \int\limits_{[Q_i^f,b]} (Y_i - Q^*)dF(Y_i) - (1+r)cQ^* =$$

$$= E_{Y_{-i}|a,b}\left[(1+r)\overline{I}_{-i} + \kappa(\overline{(Y-Q)_+})_{-i}\right] - \kappa E_{Y_i|a,b}(Y_i - Q^*)_+ - (1+r)cQ^*. \tag{6.19}$$

By Proposition 6.1 all firms j ($j \neq i$) make the first-best investment. Since the problem is symmetric, hence for all $Q_j = Q^*$, $I_j = cQ^*$ and therefore $\overline{I}_{-i} = cQ^*$ as well. Expression (6.18) depends on $\overline{I}_{-i}$ and $(\overline{(Y-Q)_+})_{-i}$ and gives us the expectation conditional on the realization of the other agent's variables. Since Y_i, $i = 1, \ldots N$ are i.i.d., then $E_{Y_{-i}|a,b}(\overline{(Y-Q)_+})_{-i} = E_{Y_{-i}|a,b}(\overline{(Y-Q^*)_+})_{-i} = E_{Y_i|a,b}(Y_i - Q^*)_+$. Therefore, (6.18) equals to zero, which finishes the proof. ■

Chapter 7

Summary of the results and conclusions

In this thesis we consider some issues related to the measurement of productivity and efficiency, and the evaluation of factors contributing to total factor productivity (TFP) growth.

The first part discusses alternative approaches to the TFP growth measurement, namely the original approach by Solow (1957), the traditional Index-Numbers and approaches adopted in Input-Output and DEA literature. We interrelate different measures, identifying the main reasons for their dissimilarities and the conditions under which the considered methods provide equivalent (or close) measures for TFP growth. The condition of optimizing behavior appears to be crucial in this respect. It lends theoretical support to the conventional Törnqvist or Fisher indices, while Malmquist index is free of the assumption of optimizing behavior. Malmquist indices, therefore, may incorporate the effect of efficiency change, unlike the conventional indices.

Input-Output framework provides indices of technical changes conceptually close to the conventional Solow Residual. However, they can be augmented to factor in both efficiency change and the terms-of-trade effect. This can be done if the observable prices are replaced by shadow prices obtained from the optimal allocation problem. Although, similar to DEA, the efficiency is interpreted as the potential for boosting the production to reach the production possibility frontier, there is an important difference in the meaning of the frontier in the two models. In DEA the

potential is determined by the observable best practice (possibly achieved by the other market participants), while in the augmented input-output model it comes from improving allocations of production factors within a multi-sectoral economy.

The two empirical applications considered in the first part of the thesis deal with the measurement of productivity growth in industrialized countries.

In the first we apply a method for estimation of TFP growth in a system of a few economies, based solely upon changes in the fundamentals of the economies. Since the economies participating in the system are linked by free trade, changes in tastes, endowments or technologies in any of them affect valuations of inputs in all economies and, therefore, influence TFP growth. Thus international trade contributes to TFP growth in each economy. TFP growth has been evaluated at shadow prices and decomposed into three terms, namely the Solow Residual, efficiency change, and the effect of change in term of trade.

The theory has been applied to estimate the TFP growth in the US, Japan and Europe (an aggregate of the UK, France and Germany) in 1985-1990. Since the system encompasses three major open economies, terms of trade are endogenous in the model. We have found that the Solow Residual corresponding to shadow prices and optimal activity levels are strongly correlated with the conventional measure of TFP growth. Japan had the highest aggregate TFP growth over the observed period. This was achieved mostly by technical change. In contrast, most of the European TFP growth was due to efficiency change.

The second application uses Malmquist indices to evaluate the TFP growth in manufacturing in selected OECD countries during the period 1970-1990. To construct the indices we apply both DEA with contemporaneous frontiers and DEA with sequential frontiers. It has been demonstrated that both methods produce highly correlated results for the total measure of TFP growth, but less correlated results for the decomposition into technical changes and efficiency changes. The sequential measure takes past information into account and reallocates temporary backwards shifts in the productivity of the best-practice countries to the efficiency change component, whilst the contemporaneous measure accounts for them as technical regress. The former is more suitable for measuring technical changes in manufacturing.

The thesis suggests a decomposition of Malmquist indices, which links the two measures of TFP growth. The new decomposition distinguishes three sources of TFP growth: technical progress, catching-up and the business cycle.

The empirical analysis has shown that most productivity increase in manufacturing in the OECD countries can be ascribed to technical progress. Five out of the six considered manufacturing sectors showed little or no catching up. Only in chemicals efficiency changes were substantial. This sector shows the strongest convergence of TFP levels. The contribution of the business cycle component of TFP growth appeared to be negative in most cases.

The second part of the thesis deals with the regulation of a natural monopoly. Related issues are the evaluation of the productivity and efficiency performance of regulated companies and the design of a regulation scheme that provides incentives for adequate performance.

Focusing on the regulation of regional distribution utilities operating under similar circumstances and able to achieve the same minimum cost, it has been shown how the traditional yardstick competition scheme can be augmented to resolve the trade-off between price and reliability of service. The suggested regulation scheme incorporates the aspect of capacity choice under uncertainty. The scheme allows the regulator to enforce the desired level of investment, corresponding to the socially-optimal level of reliability of supply, and allocate the welfare gains to the customers.

It has been proven that a scheme that does not penalize network failures, is suboptimal and leads to underinvestment. In contrast, the socially-optimal outcome can be achieved by introducing penalties for undersupply equal to the customer value of the associated losses. Then the potentially external costs of inadequate supply are internalized by the companies and hence taken into account in making their investment decisions.

In the considered model companies are exposed to the risk of a shortfall. Therefore, in order that the individual rationality constraint under demand uncertainty is met, the regulator should offer a firm an expected price which exceeds the minimum cost of installing a unit of capacity. It has been suggested how this markup, which is associated with the risk of undersupply might be quantified by a regulator when there is asymmetric information.

Bibliography

[1] Abramovitz, M. (1986) Catching up, forging ahead, and falling behind, *Journal of Economic History*, 46(2), 385-406.

[2] Ahn, B.-H., Kim, J.-C., Moon, H.-J. (1992) Disutility and constrained quality choice in self-selection problems, *Journal of Regulatory Economics*, 4(2), 159-174.

[3] Aulin-Ahmavaara, P. (1999) Effective rates of sectoral productivity change, *Economic Systems Research*, 11(4), 349-364.

[4] Barro, R.J., Sala-i-Martin, X. (1991) Convergence across states and regions, *Brookings Papers on Economic Activity*, 1, 107-158.

[5] Averch, H., Johnson, L.L. (1962) Behaviour of the firm under regulatory constraint, *American Economic Review*, 52, 1052-1069.

[6] Baumol, W.J., Klevorick A.K. (1970) Input choices and rate-of-return regulation: an overview of the discussion, *The Bell Journal of Economics and Management Science*, 1(2), 162-190.

[7] Baumol, W.J., Nelson, R.R., Wolff, E.N. (1994) *Convergence of Productivity: Cross-National Studies and Historical Evidence*, Oxford University Press, Oxford, UK.

[8] Baumol, W.J., Panzar, J.C., Willig, R.D. (1982) *Contestable Markets and the Theory of Industrial Structure*, Harcourt Brace Jonovich, New York.

[9] Bogetoft, P. (1994) Incentive-efficient productive frontiers: an agency perspective on DEA, *Management Science*, 40, 959-968.

[10] Bogetoft, P. (1995) Incentives and productivity measurement, *Production Economics*, 39, 67-81.

[11] Bogetoft, P. (1997) DEA-based yardstick competition: the optimality of best-practice regulation, *Annals of Operations Research*, 73, 277-298.

[12] Bogetoft, P. (2000) DEA and activity planning under asymmetric information, *Journal of Productivity Analysis*, 13, 7-48.

[13] de Borger, B., Kerstens, K. (2000) The Malmquist productivity index and plant capacity utilization, *Scandinavian Journal of Economics*, 102(2), 303-310.

[14] Brown, G., Johnson, M.B. (1969) Public utility pricing and output under risk, *American Economic Review*, 59, 119-128.

[15] Carree M.A., Kolmp, L., Thurik, A.R. (2000) Productivity convergence in OECD manufacturing industries, *Economic Letters,* 66, 337-345.

[16] Caves, D.W., Christensen, L.R., Diewert, W.E. (1982) The economic theory of index numbers and the measurement of input, output and productivity, *Econometrica,* 50(6), 1393-1414.

[17] Caves, D.W., Herriges, J.A., Windle, R.J. (1990) Customer demand for service reliability in the electric power industry: a synthesis of the outage cost literature, *Bulletin of Economic Research*, 42(2), 79-119.

[18] Cella, G., Pica, G. (2001) Inefficiency spill-overs in five OECD countries: an interindustry analysis, *Economic Systems Research*, 13(4), 405-416.

[19] Coate, S., Panzar, J.C. (1989) Public utility pricing and capacity choice under risk: a rational expectations approach, *Journal of Regulatory Economics*, 1(4), 305-317.

[20] Coelli, T., Psarada Rao, D.S. (2001) Implicit value shares in Malmquist TFP index numbers, CEPA Working paper No. 4/2001, University of New England, Armidale, Australia.

[21] Courville, L. (1974) Regulation and efficiency in the electric utility industry, *Bell Journal of Economics*, 5(1), 53-74.

[22] Cubbin, J., Tzanidakis, G. (1998) Regression versus data envelopment analysis for efficiency measurement: an application to the England and Wales regulated water industry, *Utilities Policy*, 7, 75-85.

[23] Diewert, W.E. (1976) Exact and superlative index numbers, *Journal of Econometrics*, 4, 115-146.

[24] Diewert, W.E. (1992) Fisher ideal output, input and productivity indexes revisited, *Journal of Productivity Analysis*, 3, 211-248.

[25] Diewert, W.E., Morrison, C. (1986) Adjusting output and productivity indexes for changes in terms of trade, *Economic Journal*, 96, 659-679.

[26] Dietzenbacher, E., Los, B. (1998) Structural decomposition techniques: sense and sensitivity, *Economic System Research*, 10(4), 307-323.

[27] Dismukes, D., Cope III, R., Mesyanzhinov, D. (1998) Capacity and economies of scale in electric power transmission, *Utilities Policy*, 7, 155-162.

[28] Dowrick, S., Nguyen, D.-T. (1989) OECD comparative economic growth 1950-1985: catch up and convergence, *American Economic Review*, 79(5), 1010-1030.

[29] Färe, R., Grabowski, R., Grosskopf, S.(1985) Technical efficiency of Philippine agriculture, *Journal of Productivity Analysis*, 17, 205-214.

[30] Färe, R., Grosskopf, S.(1992) Malmquist indexes and Fisher ideal indexes, *Economic Journal,* 102(410), 158-160.

[31] Färe, R., Grosskopf, S.(1996) *Intertemporal Production Frontiers: with Dynamic DEA*, Kluwer Academic Publishers, Boston.

[32] Färe, R., Grosskopf, S., Lindgren, B., Roos, P. (1989) Productivity development in Swedish hospitals: a Malmquist output index approach, Discussion paper No. 89-3, Southern Illinois University, Illinois.

[33] Färe, R., Grosskopf, S., Norris, M., Zhang, Z. (1994) Productivity growth, technical progress, and efficiency change in industrialized countries, *American Economic Review*, 84(1), 66-83.

[34] Farrell, M.J. (1957) The measurement of productive efficiency, *Journal of the Royal Statistical Society, Series A - General*, 120(3), 253-290.

[35] Gasmi, F., Laffont, J.-J., Sharkey,W.-W. (1999) Empirical evaluation of regulatory regimes in local telecommunications markets, *Journal of Economics and Management Strategy*, 8(1), 61-93.

[36] Gilbert, R.J., Newbery, D.M. (1988) Regulation games, Working paper No. 8879, University of California, Berkeley, California.

[37] Giulietti, M., Waddams Price, C. (2000) Incentive regulation and efficient pricing: empirical evidence, CMuR Research paper No. 00/2, University of Warwick, Warwick, UK.

[38] Gouette, C., Perelman, S. (1997) Productivity convergence in OECD service industries, *Structural changes and Economic Dynamics*, 8, 279-295.

[39] Hicks, J.R. (1935) Annual survey of economic theory: the theory of monopoly, *Econometrica*, 1, 1-20.

[40] Huettner, D., Landon, J. (1978) Electric utilities: scale economies and diseconomies, *Southern Economic Journal*, 44, 883-912.

[41] Jen, F., Tschirhart, J. (1979) Behavior of a monopoly offering interruptible service, *Bell Journal of Economics*, 10, 244-258.

[42] Jorgenson, D., Griliches, Z. (1967) The explanation of productivity change, *Review of economic Studies*, 34 (3), 249-283.

[43] Joskow, P. and R. Schmalensee (1986) Incentive regulation for electricity utilities, *Yale Journal on Regulation*, 4, 1-49.

[44] Kahn, A. (1995) *The Economics of Regulation: Principles and Institutions*, sixth printing, MIT Press, Cambridge, Massachusetts.

[45] Kittelsen, S. (1993) Stepwise DEA. Choosing variables for measuring technical efficiency in Norwegian electricity distribution. Memorandum 6/93 from the Department of Economics, University of Oslo, Norway.

[46] Laffont, J.-J., Tirole, J. (1993) *A Theory of Incentives in Procurement and Regulation*, MIT Press, Cambridge, Massachusetts.

[47] Langset, T. (2001) Quality dependent revenues - incentive regulation of quality of supply, NVE, Norway.

[48] Leibenstein, H. (1966), Allocative Efficiency vs "X-efficiency", *The American Economic Review*, 56 (3), 392-415.

[49] Littlechild, S. (1983) *Regulation of British Telecommunications*, HMSO, London, UK.

[50] Lowe, P., Fernandes, E. (1994) The growth and convergence of manufacturing productivity in industrial and newly industrialising countries, *International Journal of Production Economics*, 34, 139-149.

[51] Lyon, T. (1991) Regulation with 20-20 hindsight: "heads I win, tails you lose?", *Rand Journal of Economics*, 22(4), 581-595.

[52] McAfee, R.P., McMillan, J., Reny, P.J. (1989) Extracting the surplus in the common-value auction, Econometrica, 57(6), 1451-59.

[53] McAfee, R.P., Reny, P.J. (1992) Correlated information and mechanism design, Econometrica, 60(2), 395-421.

[54] Maudos, J., Pastor, J.-M., Serrano, L. (2000) Convergence in OECD countries: technical change, efficiency and productivity, *Applied Economics*, 32, 757-765.

[55] Mikkers, M.C., Shestalova, V. (2001) Yardstick competition and reliability of supply in public utilities, to appear in NMa Discussion papers, NMa, The Hague, The Netherlands.

[56] Negishi, T. (1960) Welfare economics and existence of an equilibrium for a competitive economy, *Metroeconomica*, 12, 92-97.

[57] Newbery, D.M. (1999) *Privatization, Restructuring, and Regulation of Network Utilities*, MIT press, Cambridge, Massachusetts.

[58] Nickerson, D.B., Reynolds, S.S. (1990) Optimal monopoly investment and capacity utilization under random demand, Working paper No. 90-003, Federal Reserve Bank of St. Louis, Missouri.

[59] Perelman, S. (1995) R&D, technological progress and efficiency change in industrial activities, *Review of Income and Wealth*, 41(3), 349-366.

[60] Peterson, W. (1979) Total factor productivity in the UK: a disaggregated analysis, in: Patterson, K.D., Schott, K.(eds), *The Measurement of Capital: Theory and Practice*, MacMillan Press, London, UK, 212-225.

[61] Pigou, A.C. (1920) *The Economics of Welfare*, MacMillan Press, London, UK.

[62] Pollitt M. (1995) *Ownership and Performance in Electric Utilities*, Oxford University Press, Oxford, UK.

[63] Ravindran, A., Phillips, D.T., Solberg, J.J. (1987) *Operations research. Principle and Practice*, second edition, Wiley Press, New York.

[64] OECD (1993) *OECD Economic Outlook*, 54, OECD, Paris, France.

[65] OECD (1995) *The OECD Input-Output Data Base*, OECD, Paris, France.

[66] OECD (1995) *Labor Force statistics, 1973-1993*, OECD, Paris, France.

[67] OECD (1996) *International Sectoral Data Base 1960-1985*, OECD, Paris, France.

[68] OECD (2001) *Structural separation in regulated industries*, report by the secretariat DAFF/CLP(2001)11, OECD, Paris, France.

[69] Saal, D., Parker, D. (2000) The impact of privatisation and regulation on the water and sewerage industry in England and Wales: a translog cost function model, ABS Working paper No. RP0103, Aston Business School, Aston, UK.

[70] Shestalova, V. (2001) General equilibrium analysis of international TFP growth rates, *Economic Systems Research*, 13(4), 391-402.

[71] Shestalova, V. (2000) Sequential Malmquist indices of productivity growth: an application to OECD industrial activities, to appear in *Journal of Productivity Analysis*.

[72] Shleifer, A. (1985) A theory of yardstick competition, *Rand Journal of Economics*, 16(3), 319-327.

[73] Sibley, D. (1988) Asymmetric information, incentives and price-cap regulation, Economics discussion paper 47, Bellcore, New Jersey.

[74] Solow, R.M. (1957) Technical change and the aggregate production function, *The Review of Economics and Statistics*, 39(3), 312-320.

[75] Spulber, D. (1988) Bargaining and regulation under asymmetric information about demand and supply, *Journal of Economic Theory*, 44, 251-268.

[76] Spulber, D. (1992) Capacity-contingent nonlinear pricing by regulated firms, *Journal of Regulatory Economics*, 4, 299-319.

[77] Statistics Bureau, Management and Coordination Agency of Japan (1995) *Statistical Yearbook*, Japan.

[78] Taskin, F., Zaim, O (1997) Catching up and innovations in high- and low-income countries, *Economic Letters*, 54, 93-100.

[79] ten Raa, T. (1995) *Linear Analysis of Competitive Economies*, Harvester Wheatsheaf, London, UK.

[80] ten Raa, T., Mohnen, P. (2002) Neoclassical growth accounting and frontier analysis: a synthesis, to appear in *Journal of Productivity Analysis.*

[81] ten Raa, T., Mohnen, P. (2001) The location of comparative advantages on the basis of fundamentals only, *Economic Systems Research*, 13(1), 93-108.

[82] ten Raa, T., Wolff, E.N. (1991) Secondary products and the measurement of productivity growth, *Regional Science and Urban Economics*, 21, 581-615.

[83] ten Raa, T., Wolff, E.N. (1996) Outsourcing of services and the productivity recovery in U.S. manufacturing in the 1980s, CentER discussion papers 9689, Tilburg University, Tilburg, The Netherlands.

[84] Tirole J. (1988) *The Theory of Industrial Organization*, MIT press, Cambridge, Massachusetts.

[85] Tulkens, H., Vanden Eeckaut, P. (1995) Non-parametric efficiency, progress and regress measure for panel data: methodological aspects, *European Journal of Operational Research*, 80, 474-499.

[86] U.S. Department of Commerce (1995), *Statistical Abstract of the US,* Economics and Statistics Administration, Bureau of CENSUS, Washington.

[87] Weber, W., Domazlicky, B. (1999) Total factor productivity growth in manufacturing: a regional approach using linear programming, *Regional Science and Urban Economics*, 29, 105-122.

[88] Weisman, D.-L. (1994) Designing carrier of last resort obligations, *Information Economics and Policy*, 6(2), 97-119.

[89] Weitzman, W. (1976) On the welfare significance of national product in a dynamic economy, *Quarterly Journal of Economics*, 90, 156-162.

[90] Wolff, E.N. (1985) Industrial composition, interindustry effects and the US Productivity slowdown, *Review of Economics and Statistics*, 67, 268-77.

[91] Wolff, E.N. (1993) Productivity growth and capital intensity on the sector and industry level: specialization among OECD countries, 1970-1988. In Silverberger, G., Soete, L. (eds.), *The Economics of Growth and Technical Change*, Edward Elgar Publishing, 185-211.

[92] Wolff, E.N. (1994) Productivity measurement within an input-output framework, *Regional Science and Urban Economics*, 24, 75-92.

Samenvatting

Dit proefschrift behandelt de samenhang tussen productiviteitsgroei en efficiency. Daarbij wordt aandacht besteed aan factoren die bijdragen aan productiviteitsgroei. De groei van de output in een economie wordt volgens de Nobelprijswinnaar Solow (1957) veroorzaakt door groei van de inputfactoren arbeid en kapitaal en technische vooruitgang; hij toont aan dat de totale factorproductiviteitsgroei (TFP-groei) gelijk is aan het verschil tussen de groei van output en de groei van input. In de literatuur is steeds gedebatteerd over de manier waarop TFP groei gemeten kan worden. In het eerste deel van dit proefschrift worden verschillende methoden om productiviteitsgroei te meten besproken. Er wordt aangetoond onder welke condities de verschillende methoden aan elkaar gelijk zijn.

De meer traditionele methoden om TFP te meten (zoals de Index Numbers methoden) en methoden zoals Data Envelopment Analysis (DEA) en Input-Output Analyse verschillen voornamelijk in de manier waarop prijzen in het model verwerkt worden. In de traditionele methoden wordt uitgegaan van de veronderstelling dat prijzen door middel van concurrentie tot stand komen, dat wil zeggen dat productie verondersteld wordt efficiënt te zijn. Dit in tegenstelling tot DEA, dat het mogelijk ook inefficiency in een bepaalde sector of onderneming te kwantificeren.

Met Malmquist-indices (de met DEA corresponderende methode om TFP groei uit te drukken) is het mogelijk om productiviteitsgroei in een bepaalde sector te ontleden in groei van efficiency en technische vooruitgang. De laatste term stemt overeen met de technische vooruitgang in het model dat door Solow ontwikkeld is. In DEA kan de productiviteit vergroot worden door in een bepaalde sector efficiënter om te gaan met de productiefactoren arbeid en kapitaal.

Input-Output Analyse is door Ten Raa en Mohnen (2001) in een algemeen evenwichtsmodel toegepast en maakt het mogelijk om verschillende sectoren in een

economie en haar internationale omgeving met elkaar te vergelijken. In hun model kan de efficiency in een bepaalde economie vergroot worden door productiefactoren anders te alloceren tussen de sectoren van een economie. Het model genereert indices met betrekking tot technische verandering. Deze indices zijn conceptueel vergelijkbaar met het Solow residual. Daarnaast kan de TFP groei in dit model worden onderscheiden in efficiency veranderingen en ruilvoet-effecten.

Zowel in DEA als in de Input-Output Analyse kan inefficiency worden geïnterpreteerd als het onbenutte potentieel te verbeteren ten opzichte van de productiemogelijkhedengrens. Er is echter een belangrijk verschil tussen de modellen in de manier waarop de grens van de productiemogelijkheden wordt bepaald. In DEA wordt de grens bepaald door de best presterende geobserveerde eenheid, terwijl in de Input-Output Analyse de grens wordt gevonden door de allocatie van productie factoren te verbeteren.

In dit proefschrift worden zowel DEA als Input-Output Analyse empirisch toegepast. Ten eerste wordt in een algemeen evenwichtsmodel om productiviteitsgroei te meten toegepast op de drie grote economieën in de periode 1985-1990, namelijk de Verenigde Staten, Japan en Europa. Omdat deze economieën door een stelsel van vrijhandel met elkaar verbonden zijn, dragen veranderingen in consumentensmaak, productiefactoren , of technologie bij aan de TFP groei. Het model is gebaseerd op de gemeten veranderingen in deze 'fundamentals' van de betrokken economieën. De TFP groei in dit model wordt ontleed in technische verandering (het Solow residual), efficiency verandering en de verandering in de ruilvoet.

De hoogste TFP groei in deze periode wordt in de Japanse economie gemeten. Deze groei wordt met name veroorzaakt door technische verandering. De groei in Europa wordt met name veroorzaakt door efficiency verbeteringen.

Het in het model gemeten Solow Residual is sterk gecorreleerd met de residuals zoals die op de conventionele manier gemeten worden.

In de tweede toepassing worden Malmquist indices gebruikt om de productiviteitgroei te meten in de industrietakken van een aantal OESO landen in de periode 1970-1990. Om de indices te kunnen construeren wordt gebruikt gemaakt van DEA met een sequentiële bepaling van de grens van de productiemogelijkheden en DEA met een gelijktijdige bepaling van de grens van de productiemogelijkheden. In het proefschrift wordt aangetoond dat de methoden een sterke correlatie kennen met betrekking tot de totale productiviteitsgroei, maar een veel minder sterke correlatie met betrekking tot de elementen waarin de totale productiviteitsgroei kan worden

onderscheiden. Bij de gelijktijdige bepaling van de grens van de productiemogelijkheden wordt een achteruitgang van de grens beschouwd als technische achteruitgang. De sequentiële bepaling houdt als het ware informatie uit het verleden vast en vat een achteruitgang van de gelijktijdige grens op als efficiency verandering. Omdat technische achteruitgang niet echt voor de hand ligt in de industrie, is de laatste methode meer geschikt om de technische vooruitgang te meten.

In dit proefschrift wordt de productiviteitsgroei ontleed in technische vooruitgang, het inhalen van een efficiency achterstand en een conjuncturele component. Door het laatste component worden beide methoden aan elkaar verbonden. De empirische analyse laat zien dat de productiviteitsgroei in de industrie in de OESO landen voornamelijk wordt veroorzaakt door technische vooruitgang. Vijf van de zes in de beschouwing betrokken sectoren toonden niet of nauwelijks sprake van convergentie van efficiency scores. Alleen in de chemische sector was er sprake van substantiële efficiency veranderingen. In deze sector convergeert het niveau van productiviteit sterk. De bijdrage van de business cycle in productiviteitsgroei lijkt in de meeste gevallen negatief.

In het eerste deel van het proefschrift wordt verondersteld dat afwijkingen van concurrentie leiden tot inefficiency. In het tweede deel van het proefschrift wordt aandacht besteed aan de mogelijkheid om in een sector waarin per definitie geen competitie mogelijk is (door het bestaan van natuurlijke monopolies) de efficiency te vergroten door het reguleren van de prijzen. Het is mogelijk om natuurlijke monopolies te prikkelen om hun efficiency te vergroten door het organiseren van artificiële competitie (zogenaamde maatstafconcurrentie).

Het in het proefschrift beschreven model gaat uit van informatie asymmetrie tussen de 'principaal' (de regulator die de prijzen vast stelt) en de 'agenten' (regionale monopolies die een aan hen gedelegeerde taken uitvoeren, onder vergelijkbare omstandigheden). De taken van de 'agenten' hebben betrekking op het leveren van netwerkdiensten. De kwaliteit van de dienstverlening is afhankelijk van onomkeerbare investeringen in de capaciteit. De investeringsbeslissing wordt genomen onder onzekerheid over de ontwikkeling van de vraag.

In het systeem van maatstafconcurrentie is de prijs die een regionaal monopolie haar klanten maximaal in rekening mag brengen niet afhankelijk van de eigen kosten van de 'agent', maar wordt hij gebaseerd op de prestatie van de andere, vergelijkbare, regionale monopolies.

Maatstafconcurrentie geeft zulke sterke prikkels om de efficiency te vergroten

door kostenverlagingen, dat de efficiency winst ten koste kan gaan van de leveringszekerheid. In het proefschrift wordt aangetoond dat als het systeem van maatstafconcurrentie wordt uitgebreid met een samenhangend systeem van boetes en beloningen, de 'agenten' een optimale kwaliteit van dienstverlening zullen leveren. Daarbij worden de welvaartswinsten die bereikt worden door efficiency winsten en kwaliteitsverbeteringen doorgegeven aan de afnemers.

In het proefschrift wordt bewezen dat als er geen of een te lage boete wordt opgelegd voor het niet leveren van diensten als gevolg van onvoldoende capaciteit, de 'agenten' in een systeem van maatstafconcurrentie onvoldoende capaciteit zullen installeren. Als de 'regulator' de boetes te hoog vast stelt, leveren de agenten ook geen optimale kwaliteit, doordat ze in dat geval op kosten van de afnemers teveel capaciteit zullen installeren.

De optimale uitkomst kan worden bereikt als de boetes voor niet levering gelijk zijn aan het verlies aan nut voor de afnemer. Daardoor wordt het externe effect van kwaliteit geïnternaliseerd in de investeringsbeslissing van de 'agent'.

In het beschreven model wordt het investeringsrisico en het risico voor niet levering gedragen door de 'agenten'. De 'agenten' verlangen een opslag op de minimum kosten voor het installeren van een eenheid capaciteit om het risico van eventuele boetes te kunnen dragen. In het proefschrift wordt beschreven op welke manier deze opslag gekwantificeerd kan worden.